

How Transformers Speak: An Interacting Multi-Particle System Perspective

Nikolas Kirschstein^{*1},
Konstantin Riedl^{†2}, Alessandro Scagliotti^{‡3,4} and Massimo Fornasier^{§3,4}

¹*Technical University of Munich, School for Computation, Information and Technology,
Department of Informatics, Munich, Germany*

²*University of Oxford, Mathematical Institute, Oxford, United Kingdom*

³*Technical University of Munich, School for Computation, Information and Technology,
Department of Mathematics, Munich, Germany*

⁴*Munich Center for Machine Learning, Munich, Germany*

Abstract

The Transformer neural network architecture is a cornerstone of many modern state-of-the-art AI systems, from large language models for text generation to image segmentation for autonomous vehicles. Still, little is known about the inner working principles of Transformers and how to interpret them. In this report, we take one step towards opening the black box with a series of empirical evaluations. First, we demonstrate that in the latent space of a Transformer model the tokens cluster over time, indicating a kind of consensus dynamics. Second, we draw a connection to clustered federated consensus-based optimisation, which affords the interpretation of tokens cooperating in groups to evolve towards a consensus point that is most relevant to the group. Our work provides stepping stones for further discoveries that benefit the explainability and trustability of Transformer-based AI applications.

Contents

1	Introduction	2
2	Building Bridges	2
2.1	Residual Neural Nets as Dynamical Systems	2
2.2	Transformer Encoder as Interacting Multi-Particle System	3
2.3	Consensus Formation in Transformers	4
3	Methodology	5
3.1	Experiment Design	5
3.2	Clustering Analysis	6
4	Results	7
4.1	Tokens	7
4.2	Attention Logits	9
5	Conclusions and Future Work	9

^{*}Email: nikolas.kirschstein@tum.de

[†]Email: konstantin.riedl@maths.ox.ac.uk

[‡]Email: scag@ma.tum.de

[§]Email: massimo.fornasier@ma.tum.de

1 Introduction

Transformer models have revolutionised many fields, such as natural language processing, speech recognition, and computer vision. The key reason, however, was their computational efficiency due to the departure from recurrence mechanisms, rather than being more principled or explainable than previous approaches. Nevertheless, their widespread use in numerous applications demands investigations into their internal working principle or “reasoning process”.

A significant rigorous attempt has been made by [Geshkovski et al. \(2024\)](#), who show that a simplified and idealised version of the Transformer obeys dynamics that closely resemble the interaction in multi-particle systems. They prove that under such dynamics, the tokens cluster and, in the time limit, converge to a single point in the Transformer’s latent space. While they exemplarily demonstrate such clustering behaviour for a single selected language model and text sample, their work lacks a comprehensive empirical evaluation.

Hence, we conduct a more comprehensive empirical study to verify this clustering behaviour. In a language model context, our results can be interpreted as tokens focusing their attention on local “leaders” in alignment with the syntactic and semantic structure of sentences. If the language task is next-token prediction, the clustering behaviour indicates the forming of “consensus” on what the multinoulli distribution over probable next tokens should be. This interpretation motivates a new conceptual connection that we draw between Transformer forward dynamics and Federated Consensus-Based Optimisation (FedCBO) by shifting the focus from tokens to attention logits.

In summary, our contributions are as follows:

1. We perform extensive clustering analyses on the latent space of real-world Transformer models to evaluate the theoretical results of [Geshkovski et al.](#) by generating detailed visualisations and employing state-of-the-art clustering methods.
2. We widen the experimental scope by also considering the widely used BERT model that violates the shared-weights assumption of the theory, as well as further architectural variations, to identify assumptions from the theory that might not be necessary.
3. We extend the existing theory by establishing a novel conceptual link between Transformer forward dynamics and FedCBO algorithms, which we also evaluate empirically by investigating the attention logits.

The remainder of the report is structured as follows: Section 2 establishes fundamental conceptual links to dynamical systems and in particular interacting multi-particle systems, Section 3 explains our methodology regarding experiment design and cluster analysis, Section 4 discusses our findings both for tokens themselves and their pairwise attention logits, and Section 5 concludes the report with future avenues to explore.

2 Building Bridges

The goal of this section is to erect several conceptual bridges, setting the stage for our empirical evaluations. Section 2.1 connects residual neural networks to dynamical systems, Section 2.2 connects Transformer encoders to interacting multi-particle systems, and Section 2.3 connects this interaction to federated consensus-based optimisation, which is the key connection that we would like to emphasise.

2.1 Residual Neural Nets as Dynamical Systems

Residual neural networks can be viewed as numerical ODE solvers in the following way: consider the hidden state $x : [0, \infty) \rightarrow \mathbb{R}^d$ as a function of time and assume the network evolves the input

$x_0 \in \mathbb{R}$ according to a first-order ODE of the form

$$\frac{d}{dt}x(t) = f(x(t), t), \quad x(0) = x_0. \quad (1)$$

When solving numerically using the forward Euler method via the first-order Taylor expansion

$$x(t + \tau) = x(t) + \tau f(x, t) + \mathcal{O}(\tau^2) \quad (2)$$

and dropping all higher-order terms, we obtain the iterative algorithm

$$x^{(\ell+1)} \leftarrow x^{(\ell)} + \tau f(x^{(\ell)}, t^{(\ell)}), \quad x^{(0)} \leftarrow x_0, \quad t^{(0)} \leftarrow 0 \quad (3)$$

which for sufficiently small τ approximates $x^{(\ell)} \approx x(\ell\tau)$. However, equation (3) is nothing else but a feed-forward residual neural network with generic layer function f that can encode any architecture by virtue of time-dependence.

2.2 Transformer Encoder as Interacting Multi-Particle System

2.2.1 Linear Algebra Formulation

Transformer encoders (Vaswani et al., 2017) are feed-forward residual neural networks consisting of an alternation of multi-head self-attention layers and position-wise affine layers, each of which are followed by layer normalisation and a residual connection. Given an input $X \in \mathbb{R}^{d \times d}$ where rows are abstract tokens, the processing, parametrised by θ_1, θ_2 , can be written as

$$\text{TransformerBlock} : \begin{cases} Y \leftarrow X + \text{LayerNorm}(\text{MultiHeadSelfAttention}(X; \Theta_1)) \\ Z \leftarrow Y + \text{LayerNorm}(\text{Dense}(Y; \Theta_2)) \end{cases}. \quad (4)$$

Dense layer is a synonym for affine transformation and layer normalisation is row-wise Z-score standardisation followed by a learnt element-wise scaling and shifting. The centrepiece of a Transformer encoder consists in multi-head self-attention, which integrates the output of several attention heads. Each head $h = 1, \dots, H$ is computed via self-attention on its input as

$$\text{Attention}(X; Q_h, K_h, V_h) := \text{softmax} \left(\beta X Q_h K_h^\top X \right) X V_h \quad (5)$$

where $Q_h, K_h \in \mathbb{R}^{d \times d'}$ and $V_h \in \mathbb{R}^{d \times d}$ are the learnable parameters of head h , and $d' \mid d$

$$\begin{aligned} \text{MultiHeadSelfAttention}(X; (Q_h, K_h, V_h)_{h=1}^H) &:= \\ &= [\text{Attention}(X; Q_1, K_1, V_1) \quad \dots \quad \text{Attention}(X; Q_H, K_H, V_H)] \underbrace{\begin{bmatrix} W_1 \\ \vdots \\ W_H \end{bmatrix}}_{\substack{\text{denoted as } W^O \\ \text{in other works}}} \end{aligned} \quad (6)$$

Attention! Note that we intentionally deviate from the usual notation by Vaswani et al. where Q, K, V refer to the products XQ, XK, XV rather than the parameters themselves, which in turn are normally denoted as W^Q, W^K, W^V . The reason for our deviation is notational convenience and conceptual clarity, as will become apparent in the upcoming sections.

2.2.2 Dynamical System Formulation

The multi-particle point of view on Transformers was first introduced by [Lu et al. \(2019\)](#) and substantially extended by [Geshkovski et al. \(2023, 2024\)](#). For their theoretical results, they make the following simplifying assumptions that remove some of the bells and whistles of a full Transformer.

Assumption 1. *There are no dense layers, i.e., $\text{Dense} := 0$.*

Assumption 2. *There is only a single attention head, i.e., $H = 1, d' = d$.*

Assumption 3. *Layer normalisation without scale and shift behaves similar to a projection onto the tangent space of an axis-aligned ellipsoid and that ellipsoid is the unit sphere, i.e., no learnt scale and shift are re-introduced.*

Under these assumptions, Transformer encoders can be seen as a flow map of particles $(x_i)_{i=1}^n$ on the d -dimensional unit sphere $\mathbb{S}^{d-1}$. Considering each x_i individually as a column vector from now on, the flow map is given through the per-particle dynamics

$$\frac{d}{dt}x_i(t) = P_{x_i(t)} \left[\frac{1}{Z_i(t)} \sum_{j=1}^N \underbrace{\exp(\beta \langle Q(t)x_i(t), K(t)x_j(t) \rangle)}_{=:w_{ij}} V(t)x_j(t) \right] \quad (7)$$

with softmax normalisation factor $Z_i(t) := \sum_{j=1}^N w_{ij}$ and orthogonal projection $P_x y := y - \langle x, y \rangle x$ onto the tangent space $T_x \mathbb{S}^{d-1}$. The projection is a reasonable proxy for layer normalisation as the latter also forces the dynamics to remain on the unit sphere: if μ and σ are the mean and standard deviation of the coordinates of token $x_i(t)$, then

$$\left\| \frac{x - \mu}{\sigma} \right\|_2 = \frac{1}{\sigma} \underbrace{\|x - \mu\|_2}_{\sigma} = 1.$$

The main result of [Geshkovski et al.](#) is that, under the two additional assumptions stated below, a single cluster emerges for $t \rightarrow \infty$ in sufficiently high-dimensional latent spaces.

Assumption 4 (Time-independence). *The weights $Q(t) = Q, K(t) = K$ are time-independent, i.e., shared across layers.*

Assumption 5 (Identity value mapping). *The value matrices $V(t) = I$ are identity mappings.*

Theorem 1 ([Geshkovski et al.](#), Theorem 4.1). *Let $d \geq N$. Let $(x_i)_{i=1}^N$ be the solution of an initial value problem where the ODE is equation (7) restricted by Assumption 4, and the initial condition is a token sequence distributed uniformly at random. Then, there almost surely exist $x^* \in \mathbb{S}^{d-1}$ and $C, \lambda > 0$ such that*

$$\|x_i(t) - x^*\| \leq C \exp(-\lambda t) \quad \text{for all } i = 1, \dots, N \text{ and } t \geq 0 \quad (8)$$

2.3 Consensus Formation in Transformers

The transformer dynamics (7) are reminiscent of consensus-based optimisation (CBO) algorithms (see [Pinnau et al., 2017](#); [Carrillo et al., 2018](#); [Fornasier et al., 2024](#)). Specifically, [Carrillo et al. \(2024\)](#) introduce *federated* CBO (FedCBO), motivated by decentralised clustered federated learning. There are N agents and K objective functions. Each agent i is assigned an objective function $\mathcal{E}_i$ out of the K that it aims to minimise. From this, [Carrillo et al.](#) construct a mean-field system whose discretisation results in a practicable optimisation algorithm. At every time step t , it updates agent i as

$$x_i \leftarrow (1 - \lambda)x_i + \lambda m_i = x_i - \lambda(x_i - m_i) + \text{noise}, \quad (9)$$

where m_i is the *consensus point* of the cluster particle i belongs to. The consensus point is a weighted average of all particles s.t. particles with lower error contribute more, as they might be closer to a local minimum, while contributions of particles with very different error function are downweighted:

$$m_i \leftarrow \frac{1}{\sum_{j=1}^N w_{i,j}} \sum_{j=1}^N w_{i,j} x_j \quad \text{where} \quad w_{i,j} := w_i(x_j) := \exp(-\alpha \mathcal{E}_i(x_j)) \quad (10)$$

We now explain the marriage of FedCBO with the transformer dynamics. First, the exponential moving average update in (9) only differs from an additive residual update by a factor of $(1 - \lambda)$. Hence, the iterative update can again be seen as Euler step discretisation of an ODE whose right-hand side is given by the consensus drift term $\lambda(m_i - x_i)$. Applying the projection P_{x_i} to it would result in just λm_i due to $P_{x_i} x_i = 0$. Consequently, the FedCBO update (9) precisely matches Equation (7) if we allow for the following additional assumptions on top of Assumptions 1-5 from Section 2.2.2.

Assumption 6 (No layer normalisation). *The projection $P_{x_i(t)}$ is removed.*

Assumption 7 (Additive-noise time-dependence). *The attention logits only depend on time via additive noise, i.e., for some constant $E_i \in \mathbb{R}^d$ and zero-centred noise function n_i , we have*

$$K^\top Q x_i(t) = E_i + n_i(t). \quad (11)$$

Based on these assumptions, we can complete the bridge to Transformers by viewing the argument to the softmax as a FedCBO error function.

Definition 1. *We define particle i 's error function as the following time-independent proxy of the negative attention logits:*

$$\mathcal{E}_i(x) := -\beta \langle E_i, x \rangle \stackrel{(11)}{\approx} -\beta \langle K^\top Q x_i(t), x \rangle = -\beta \langle Q x_i(t), K x \rangle \quad (12)$$

Intuitively, $\mathcal{E}_i(x)$ checks how important x is for x_i . Now, if the error functions $\mathcal{E}_i$ form $K < N$ clusters rather than being N fully distinct entities, the synergy effects lead to the FedCBO algorithm being effective and the interpretation of grouped consensus formation is valid. This is the central theoretical aspect out of those that we empirically investigate with our methodology, which we describe next.

3 Methodology

To empirically study the bridges built in Section 2, we conduct numerical experiments to investigate our two main objects of interest: the tokens and the attention logits, alongside with their respective clustering behaviour. Section 3.1 outlines the real-world datasets and Transformer model variations we consider, and Section 3.2 explains how we perform clustering analysis. Our methodology is fully reproducible and the open-source code is publicly available at

<https://github.com/nkirschi/How-Transformers-Speak>

3.1 Experiment Design

To investigate our objects of interest under different scenarios, we consider a total of four independent variables. They amount to dataset and model architecture choices and are summarised in Table 1. In the following, we explain them in more detail:

variable	choices
data source	sentence-level (wikitext) paragraph-level (imdb) random tokens (uniform)
time-dependence	yes (BERT) no (ALBERT)
dense layers	yes (practice) no (theory)
parameters	pre-trained randomly initialised

Table 1: Overview of independent variables in our experiments.

Datasets. We employ three different text datasets. The **wikitext**¹ dataset contains sentences from verified articles on Wikipedia and has been used by Geshkovski et al.. For a paragraph-level dataset, we use the curated **imdb**² dataset. In addition to these two real-world datasets, we also add a synthetic dataset **uniform** to the comparison which contains token sequences drawn i.i.d. uniformly at random from the vocabulary, matching the premise of Theorem 1. For each dataset, we only consider samples consisting of at least 300 tokens and randomly draw ten of those.

Models. Due to limited computational resources and time constraints, we could not conduct experiments on currently available large language models like Llama 2 (Touvron et al., 2023). Instead, we use the older but still very common BERT_{LARGE}³ (Devlin et al., 2019) model as well as ALBERT_{XLARGE}⁴ (Lan et al., 2020) which only differs from the former by weight sharing, i.e., all layers share the same parameters. Both models have 24 layers, 16 attention heads, and latent space dimensionality of 1024 and 2048, respectively. As ALBERT uses the same parameters for all layers, we are able to artificially “extend” the dynamics beyond 24 layers. Since the theory only considers a single attention head, we always choose a random head of the 16 to analyse.

Variations. Comparing the two models allows us to see the empirical effect of time-dependence of parameters, which is not yet covered by the theory. Another missing element in the theory are the dense layers in between attention layers, so we also compare the difference in dynamics when adding/removing them from the architecture. Finally, we conduct all experiments with both with the pre-trained model parameters as well as randomly initialised parameters, according to the model’s individual initialisation scheme.

3.2 Clustering Analysis

Since centroid-based clustering methods in the spirit of k -means require the a priori specification of the number of clusters as a hyperparameter, we employ a density-based clustering algorithm that dynamically determines the number of clusters. The current state of the art in this territory is HDBSCAN⁵ (Campello et al., 2013), a hierarchical version of the established DBSCAN algorithm (Ester et al., 1996). While DBSCAN assumes a globally uniform density of clusters that needs to be specified via hyperparameters, HDBSCAN abstracts away from this by constructing a hierarchy of heterogeneous-density clusters and finds an optimal cut in this hierarchy with respect to a sophisticated cluster stability measure.

¹<https://huggingface.co/datasets/Salesforce/wikitext>

²<https://huggingface.co/datasets/stanfordnlp/imdb>

³<https://huggingface.co/google-bert/bert-large-uncased>

⁴<https://huggingface.co/albert/albert-xlarge-v2>

⁵Hierarchical Density-Based Spatial Clustering of Applications with Noise

Due to the curse of dimensionality implied by only hundreds of tokens in the high-dimensional latent space of the (AL)BERT models, we cannot apply density-based methods directly in the original space as the notion of density becomes meaningless. The authors of the HDBSCAN Python library suggest a dimensionality of at most 50. Hence, we first apply standard Principal Component Analysis (PCA) to reduce the dimensionality to 32. It is on this PCA-reduced space where we apply HDBSCAN. Despite the abstraction it affords, there still is one hyperparameter that needs to be specified which defines the minimum required number of points for a grouping to be considered a cluster. We set this rather conservatively to 3 due to the low number of tokens that we seek to cluster, which is on the order of a few hundred.

The quality of thus obtained clustering assignment is assessed on the following standard metrics:

- *Outlier rate*: The fraction of samples considered outliers and thus not assigned to any cluster by the algorithm. It takes values in $[0, 1]$ and a lower outlier rate suggests that the obtained clustering is more meaningful, while a high outlier rate hints at there not being any clusters at all.
- *Silhouette score*: The fraction $\frac{d_{\text{ext}} - d_{\text{int}}}{\max\{d_{\text{ext}}, d_{\text{int}}\}}$, where d_{ext} is the average inter-cluster (external) distance and d_{int} is the average intra-cluster (internal) distance to the closest cluster. It takes values in $[-1, 1]$ and a higher value indicates better clustering.
- *DBCV score*: The fraction $\frac{\rho_{\text{ext}} - \rho_{\text{int}}}{\max\{\rho_{\text{ext}}, \rho_{\text{int}}\}}$, which is analogous to the Silhouette score but considers inter- and intra-cluster *densities* $\rho_{\text{ext}}, \rho_{\text{int}}$ instead of Euclidean distances. It takes values in $[-1, 1]$ and a higher value indicates better clustering. This is the evaluation metric proposed by the inventors of HDBSCAN (see Moulavi et al., 2014).

4 Results

This section presents the results obtained with our methodology described in Section 3. It is divided into two parts, one for each of object of interest: Section 4.1 discussed the results for tokens, and Section 4.2 analogously for the attention logits. We only present selected representative visualisations and refer to the code repository⁶ for the complete collection of all 2712 plots that we generated for the 24 independent variable combinations in Table 1 and 10 input samples per combination. Every plot we include in this report contains in its title a descriptive reference to the corresponding directory in the repository.

4.1 Tokens

Since most of the independent variable combinations outlined in Table 1 exhibit similar results, we discuss one representative case in detail and then point out the differences to other combinations. This case is `imdb_albert-xlarge-v2_params-trained-dense-on-layers-48`, where we use ALBERT to stay closer to the theory in terms of time-independence, but leave the pre-trained parameters and dense layers unmodified to retain real-world relevance.

A first attempt at evaluating clustering behaviour (and the only one Geshkovski et al. employ) lies in considering how the pairwise cosine similarity of token representations in the hidden space evolves with the layers. Figure 1 visualises the cosine similarity distribution after every layer as normalised histograms with mean and confidence interval based on a sample of 10 inputs, and Figure 2 does the same for the extended ALBERT up to layer 48. We clearly see a distribution shift towards a cosine similarity of 1 with deeper layers, which implies that tokens are becoming more and more similar and may eventually form a single if evolved for many more layers. However, these histograms do not provide any insights on a potential intermediate cluster structure that might form much earlier.

⁶<https://github.com/nkirschi/How-Transformers-Speak/tree/main/visualisation>

A better way of visualising the cosine similarities is to plot them directly as heatmaps (cp. Figures 3 and 4), whereby we lose much less detail than in the histograms. In this vein, we are able to observe that the pairwise similarities exhibit an increasing amount of grid-like structure, which indeed hints at intermediate clusters forming. This starts after the first couple of layers already and becomes quite pronounced towards the final actual layers, without the need to artificially extend the dynamics for another 24 steps. In accordance with the histograms, we also observe once again that the overall “similarity heat” is constantly increasing over time.

Motivated by the insights from the heatmaps, we perform HDBSCAN clustering on the token representations as described in Section 3.2. The resulting clustering metrics are shown in Figure 5. The outlier rate is generally low and the Silhouette and DBSC scores increase with the layers, indicating an increasing clustering quality. The reported number of clusters is rather low on the order of tens but, however, quite fluctuating and thus inconclusive. One would expect a more directional decrease over time. Instead of merely looking at this flat summary metric, we plot the actual cluster sizes per layer as bar charts in Figures 6 and 7. They indicate that there are only very few dominating clusters and the remaining ones are rather small, which in part explains the observed variance in the summary metric. Especially after layer 24, the clustering becomes very pronounced.

We complement the rather technical clustering evaluation with a direct visualisation via t-distributed Stochastic Neighborhood Embeddings (t-SNE) in two dimensions. Figures 8 and 9 show the tokens’ t-SNE embeddings as points on the plane, coloured in accordance with their cluster size rank from Figures 6 and 7, but also including outlier points in yellow. The cluster structures become apparent, but we note that due to the drastic dimensionality reduction from 1024/2048 to 2 leads to a rather crude visualisation. It is aesthetic to look at, though.

The results for `imdb_albert-xlarge-v2_params-trained-dense-on-layers-48` are quite encouraging. Now we describe how they compare to other independent variable combinations (cp. Table 1). For these discussions, we do not include plots in this report but refer the reader to our repository for visualisations.

- First off, the results are very consistent across the datasets we consider. On `wikitext` they are almost identical to `imdb`. With tokens drawn from the `uniform` distribution, the clustering effect is very extreme and takes place much quicker. This makes sense given that the theory (see Theorem 1) assumes precisely such uniformly distributed inputs for the clustering result.
- Similarly, the clustering effect also tends to be more pronounced when the model parameters are randomly initialised. This is particularly interesting since the theory, so far, needs to make strong assumptions on the parameters. The clear empirical result suggests that those assumptions are not necessary for the clustering behaviour to unfold.
- Removing the dense layers from the architecture does not appear to have a noteworthy effect on the results which makes sense as they only act token-wise and do not introduce any interaction between tokens.
- The only case where the clustering effects are much weaker is when using BERT instead of ALBERT. One might think that this implies the influence of time-dependence is too strong for the time-independence assumption to be able to be dropped. However, looking at the results for uniform random tokens in this case shows that the clustering effect is still there, but unfolding much much slower. Hence, we conclude that for time-dependent parameters and real-world data, the dynamics need to run for a significantly larger number of layers than when time-independence holds.

Overall, our results suggest that the strong assumptions on the parameter matrices $Q(t)$, $K(t)$, $V(t)$ as well as the initial token distribution are not necessary for clustering to take place, but it happens much slower the further away from the assumptions we move.

4.2 Attention Logits

We now perform an empirical evaluation analogous to Section 4.1 but on our second object of interest: the pre-softmax attention logits, which relate to the error function from Definition 1. We will again use `imdb_albert-xlarge-v2_params-trained_dense-on_layers-48` as an example and observe very similar clustering behaviour, which indicates that our FedCBO connection is meaningful.

Skipping the discussion of histograms, we turn straight to the heatmap visualisations in Figures 10 and 11. Since the logits are not normalised as opposed to the cosine similarity, the domain is no longer $[-1, 1]$ but differs between layers and scenarios. Also, note that as opposed to cosine similarity matrices, the attention logit matrices need not be symmetric. Note that each row in the heatmaps corresponds to the error function of a token. The grid-like structure is appearing once again, and highly similar to that in Figures 3 and 4.

When performing HDBSCAN clustering on the rows of the attention logit matrices, the resulting metrics (cp. Figure 12) are also very similar to the ones for the token cosine similarities, with the same increase in cluster quality but inconclusiveness of cluster number. The cluster size bar charts in Figures 13 and 14 analogously show that there is only few dominant clusters, and the t-SNE embeddings in Figures 15 and 16 are again nice to marvel at.

For the comparison to the results of other independent variable combinations, we refer to the discussion in the previous Section 4.1 and the plots in the repository, as the differences and overall observations are largely analogous.

5 Conclusions and Future Work

In this report, we established crucial conceptual bridges between Transformer neural networks and interactive multi-particle system dynamics. In particular, the Transformer forward pass dynamics can be interpreted as negotiating consensus between tokens in the spirit of federated consensus-based optimisation. We then empirically evaluated the theoretical cluster formation with extensive experiments on real-world datasets and models with different variations. Our results indicate that this clustering indeed takes place, across different settings, which validates the federated consensus interpretation. Furthermore, neither the assumptions of time-independent model parameters nor absence of dense layers nor uniformly distributed input sequences seem to be necessary for the clustering behaviour.

Our work motivates a myriad of future research directions. On the theory side, the mathematical machinery needs to be advanced to require less and much weaker assumptions than is presently the case. For the clustering statements, the time-independence of model parameters, the restriction to a single attention head, and the disregard of dense layers should be challenged. For the FedCBO connection, the disregard of layer normalisation and the additive-noise time-dependence assumption deserve scrutiny. On the empirical side, we suggest to perform experiments on a recent very deep large language model like Llama 2 that has many more layers and a higher latent dimensionality since our observation suggest that in this regime the clustering may become apparent without artificial layer extension.

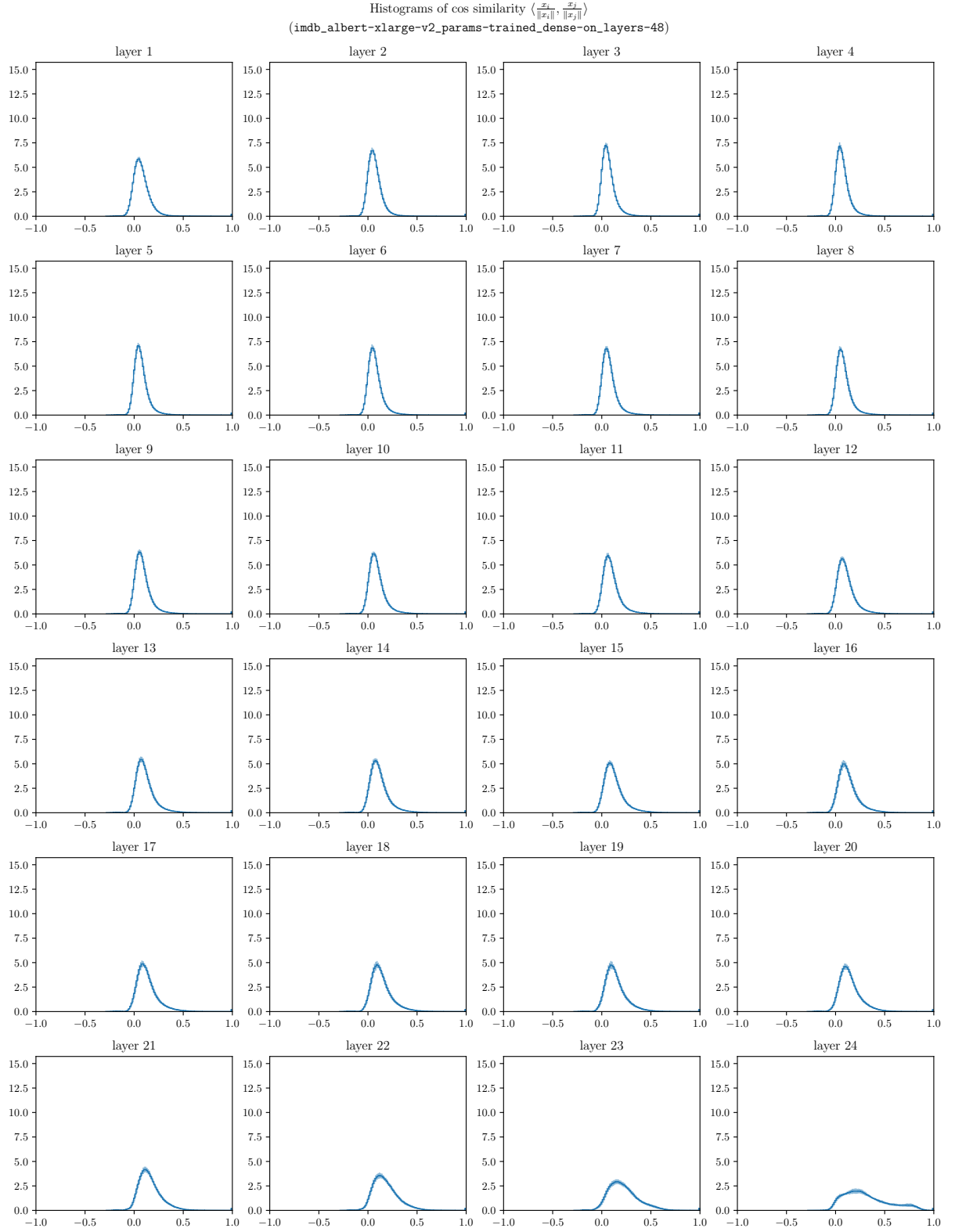


Figure 1: Token cosine similarity histograms for layers 1 to 24.

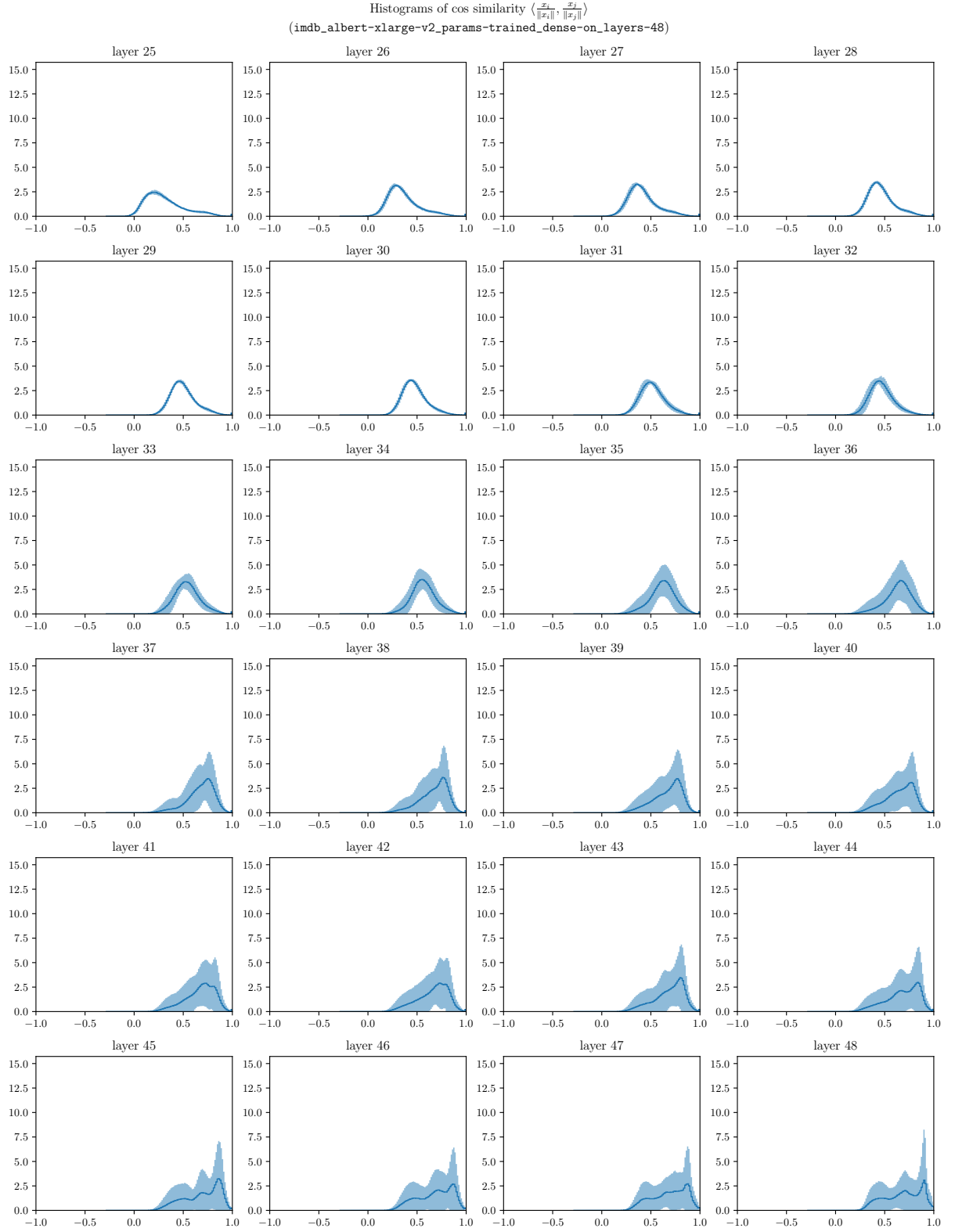


Figure 2: Token cosine similarity histograms for layers 25 to 48.

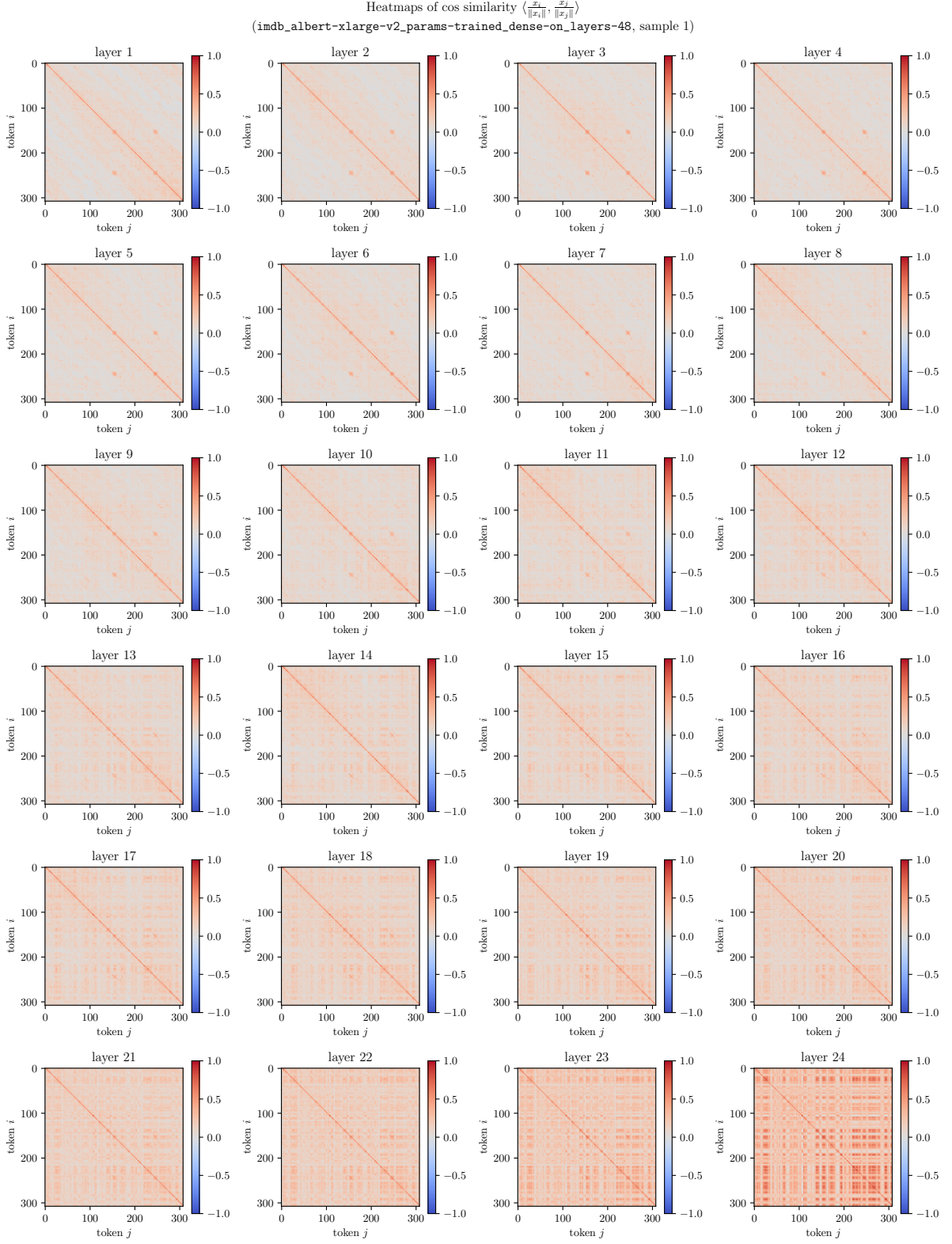


Figure 3: Token cosine similarity heatmaps for layers 1 to 24.

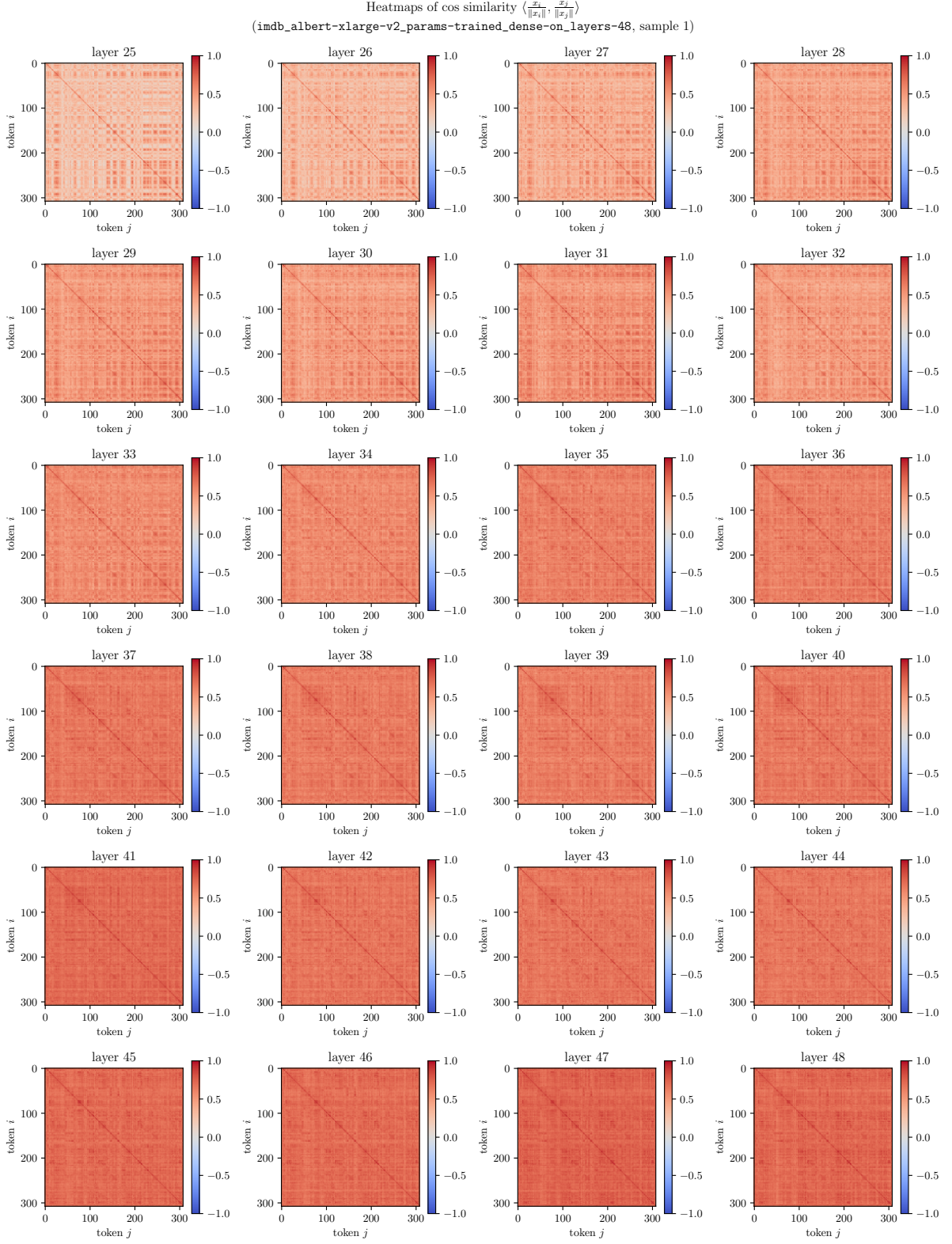


Figure 4: Token cosine similarity heatmaps for layers 25 to 48.

HDBSCAN cluster evaluation of tokens x_i
(imdb_albert-xlarge-v2_params-trained_dense-on_layers-48, sample 1)

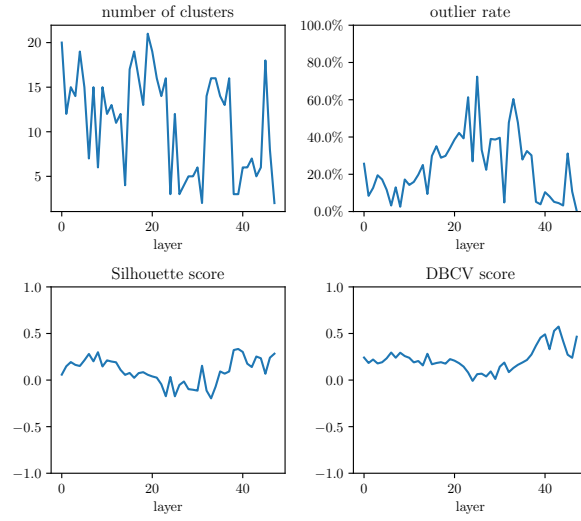


Figure 5: Token clustering metrics across layers.

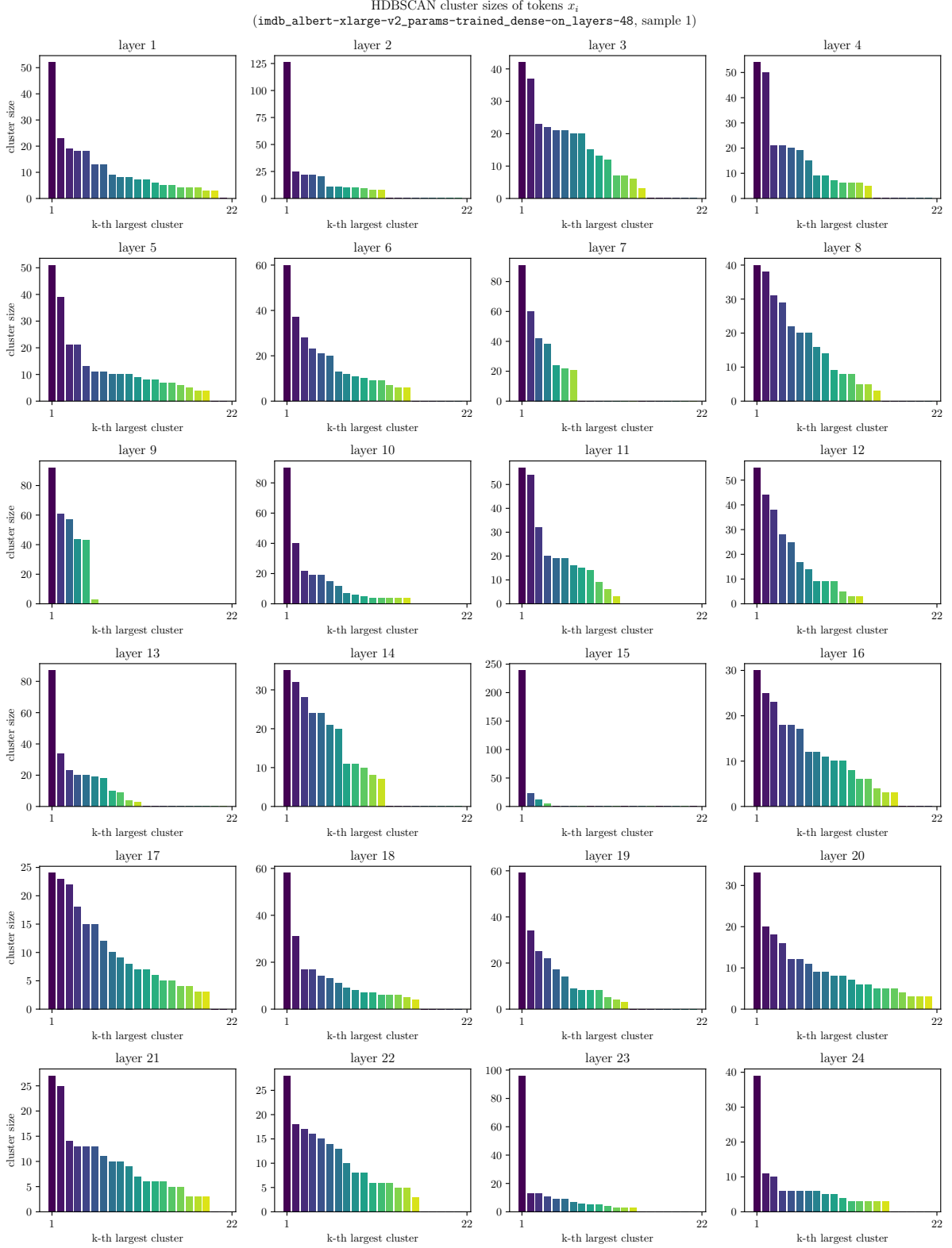


Figure 6: Token cluster sizes for layers 1 to 24. Outliers are not included.

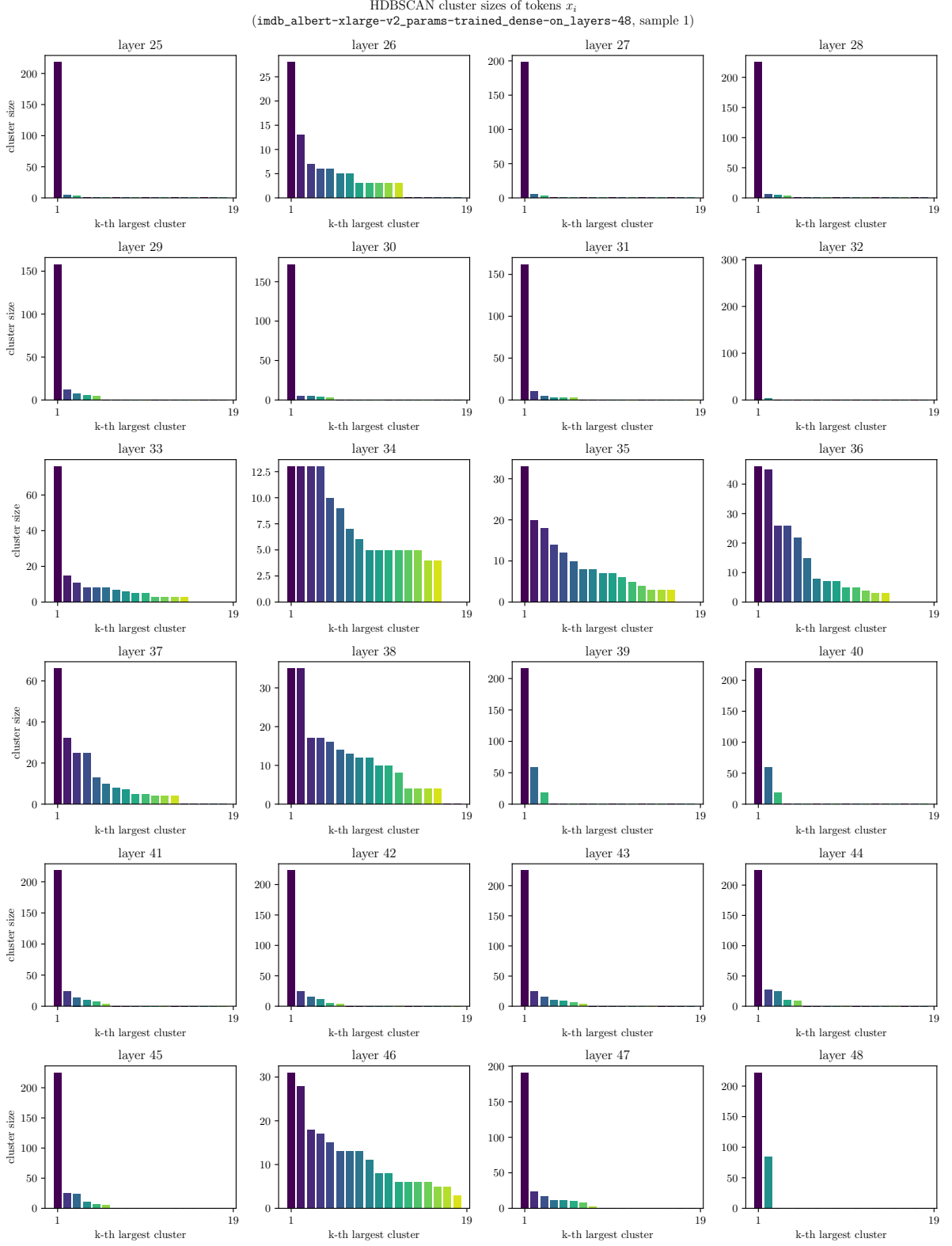


Figure 7: Token cluster sizes for layers 25 to 48. Outliers are not included.

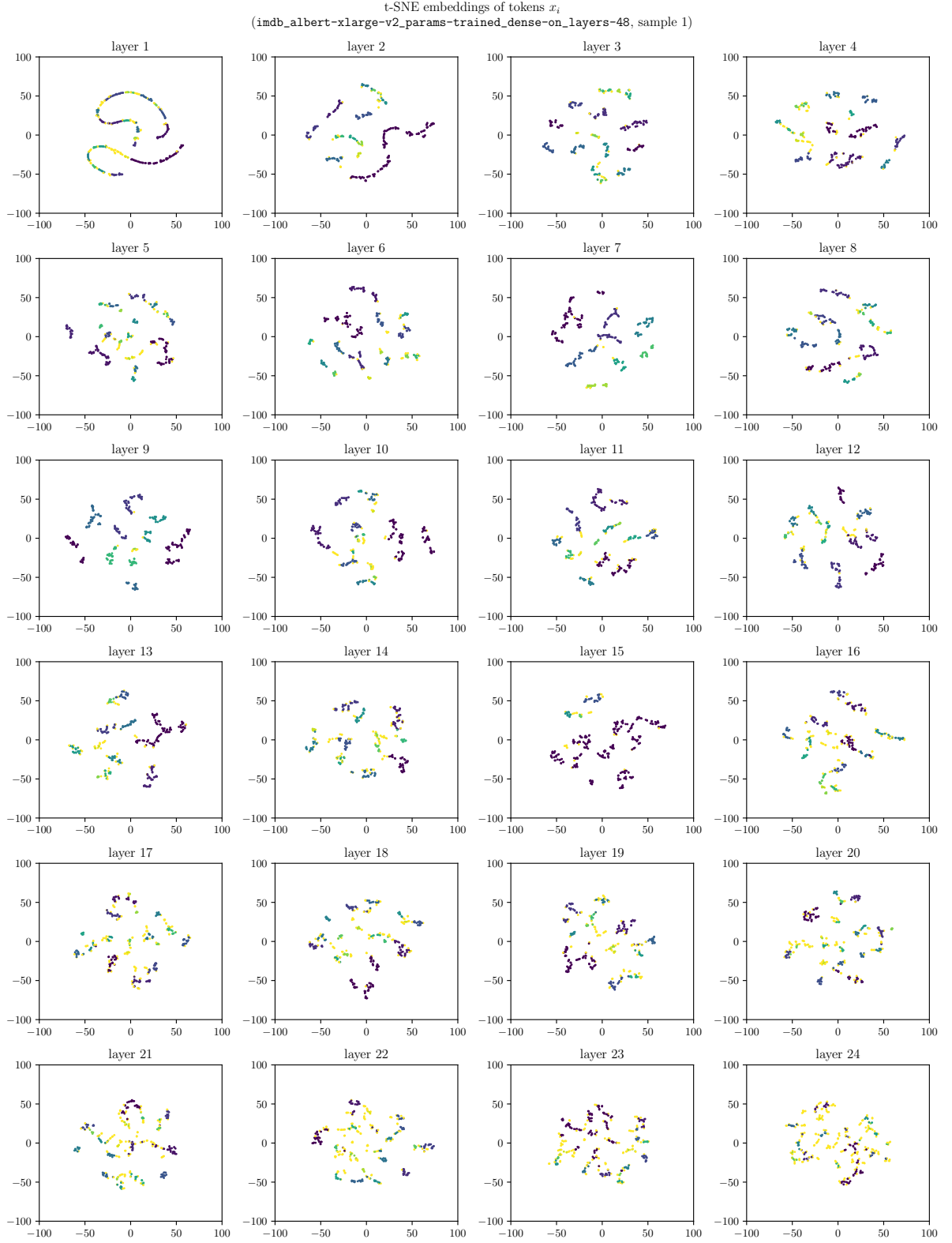


Figure 8: Token t-SNE embeddings for layers 1 to 24. Outliers are shown in yellow.

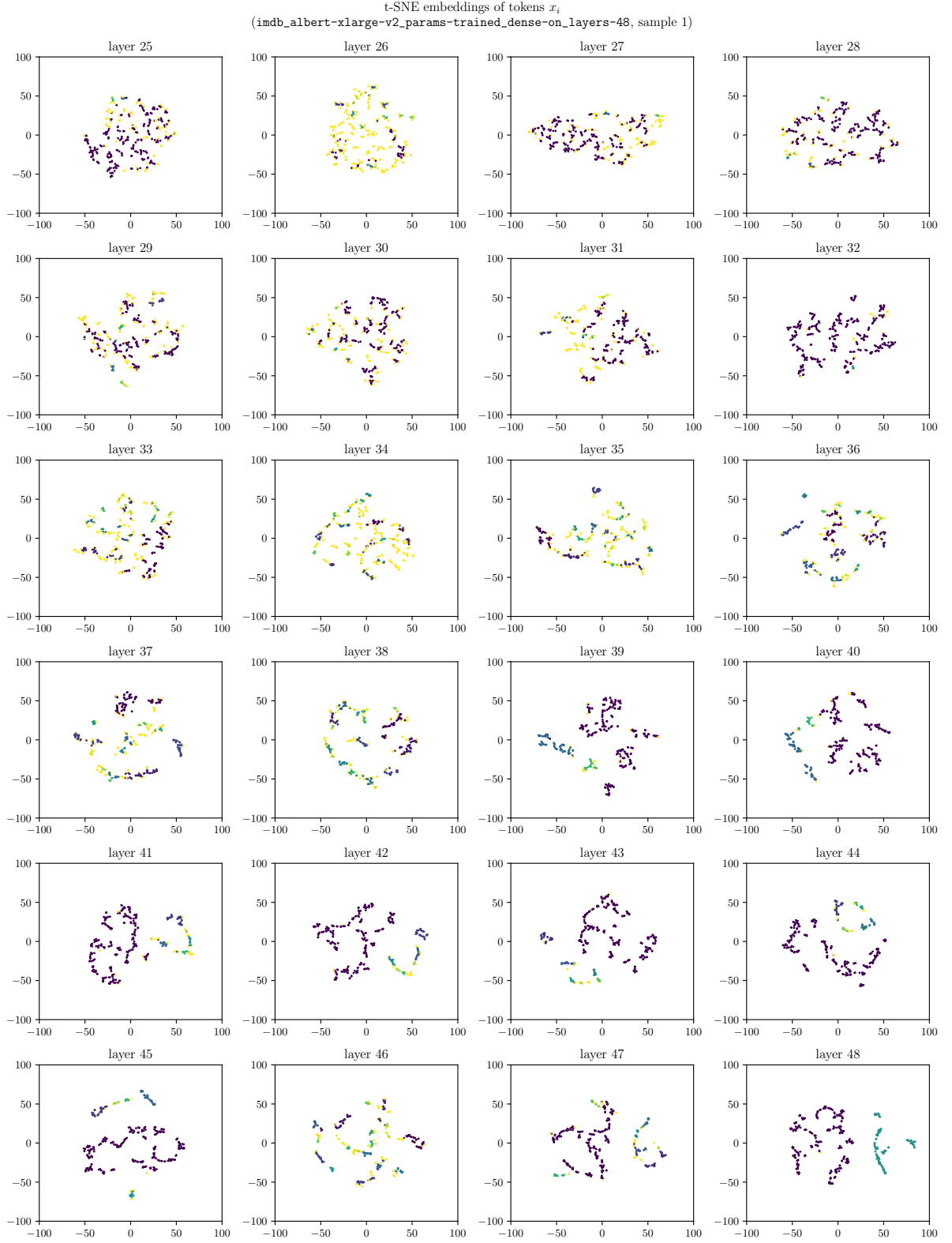


Figure 9: Token t-SNE embeddings for layers 25 to 48. Outliers are shown in yellow.

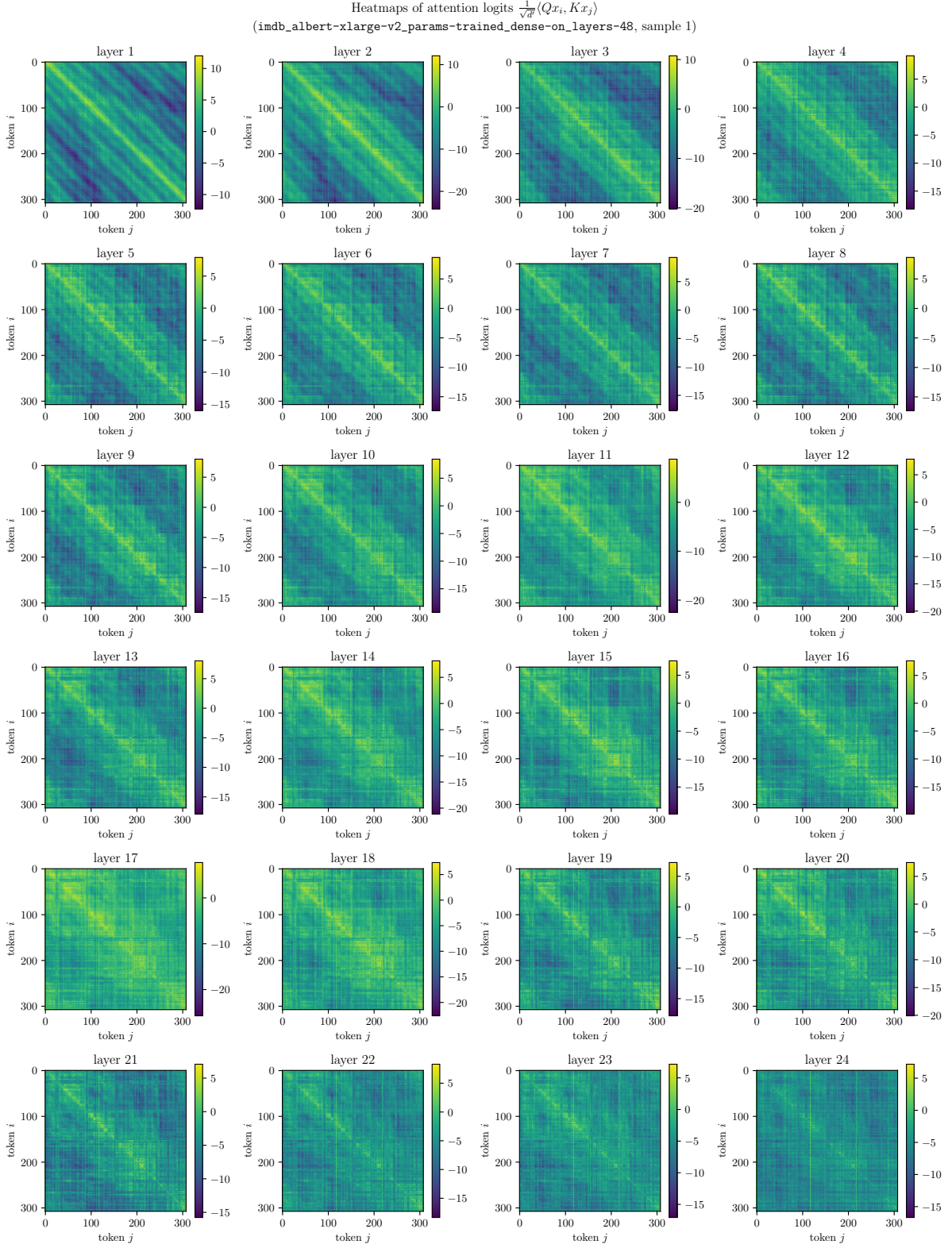


Figure 10: Attention logit heatmaps for layers 1 to 24.

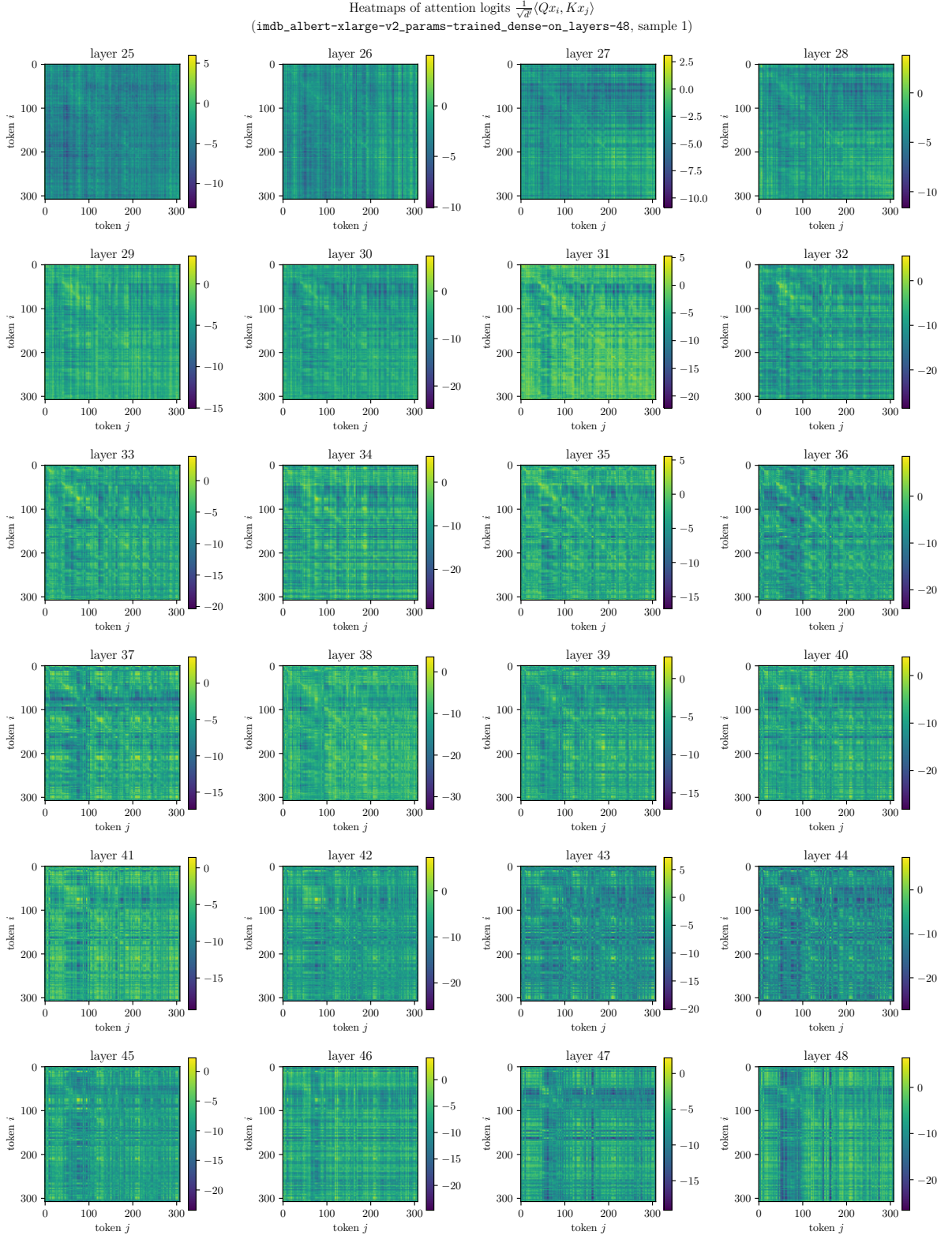


Figure 11: Attention logit heatmaps for layers 25 to 48.

HDBSCAN cluster evaluation of error functions $\mathcal{E}_i$
(imdb_albert-xlarge-v2_params-trained_dense-on_layers-48, sample 1)

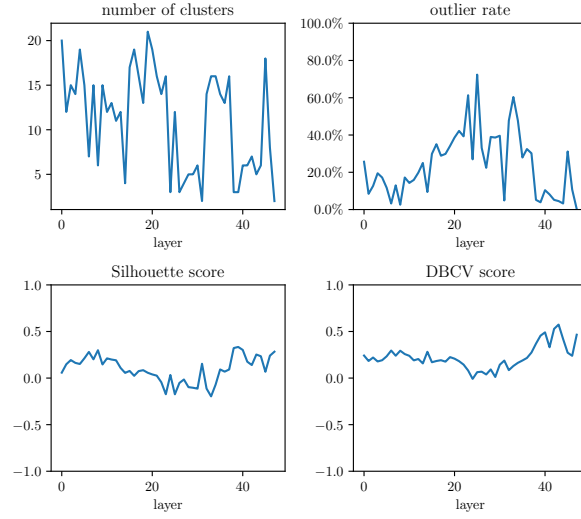


Figure 12: Error function clustering metrics across layers.

HDBSCAN cluster sizes of error functions $\mathcal{E}_i$
(imdb_albert-xlarge-v2_params-trained_dense-on_layers-48, sample 1)

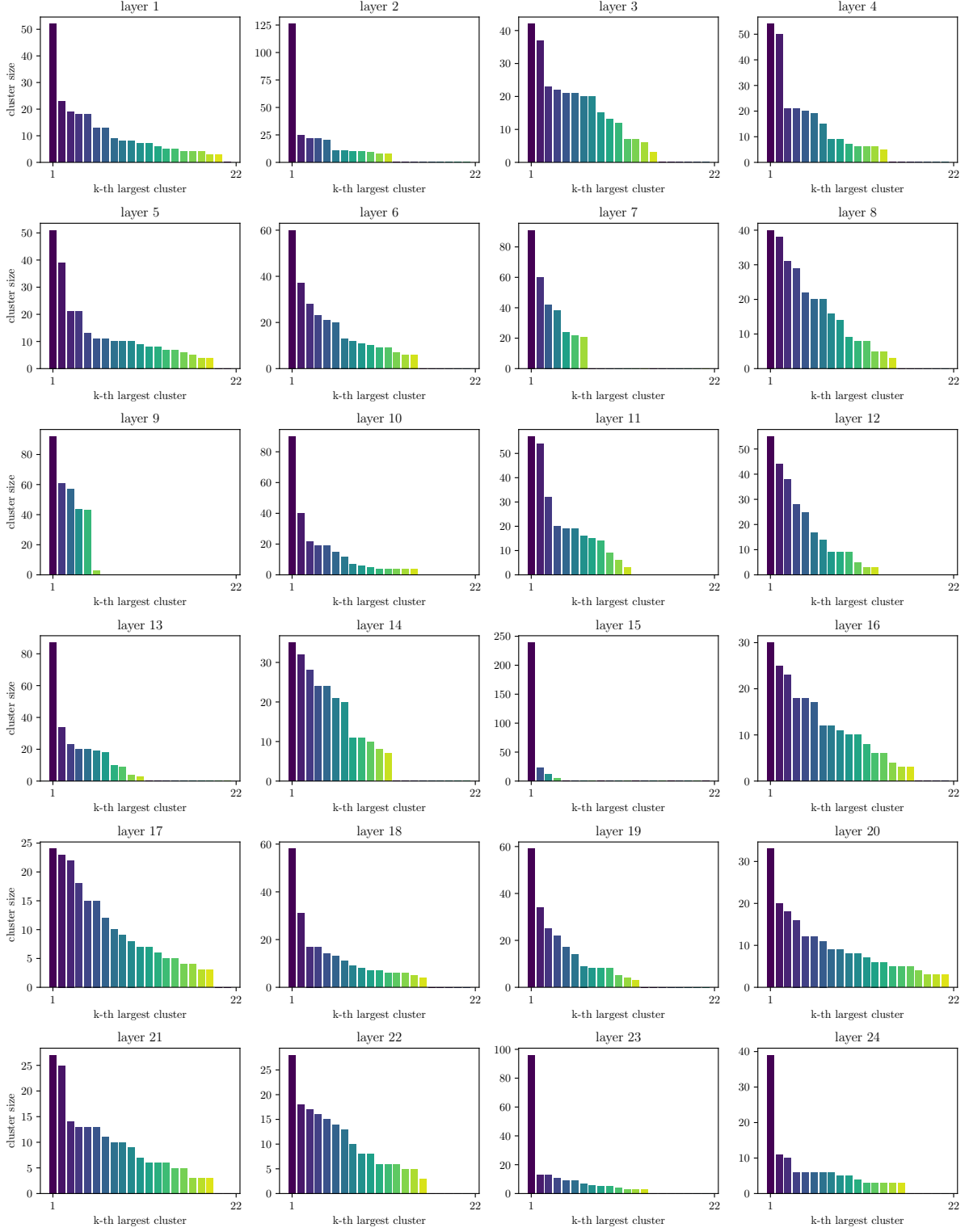


Figure 13: Error function cluster sizes for layers 1 to 24. Outliers are not included.

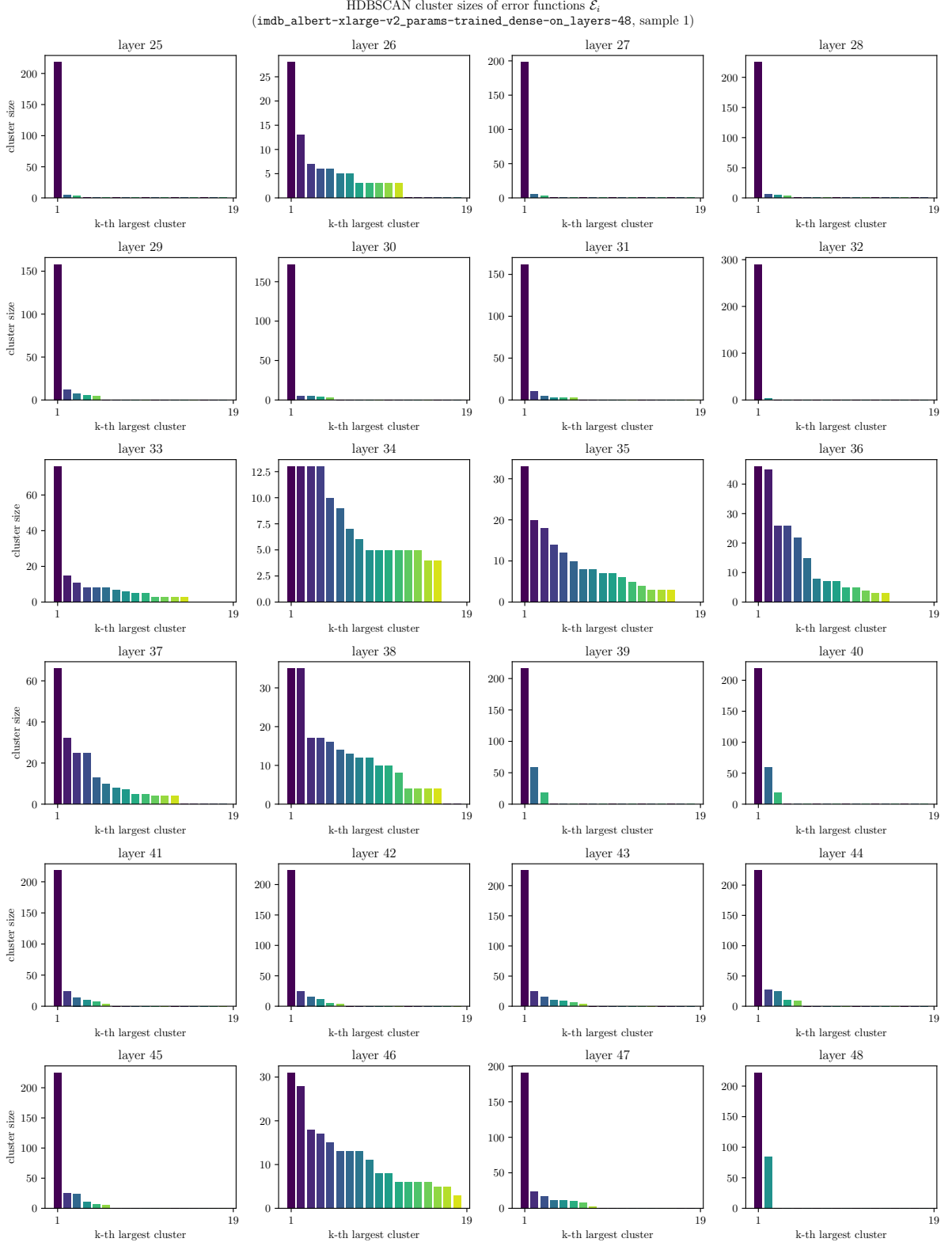


Figure 14: Error function row cluster sizes for layers 25 to 48. Outliers are not included.

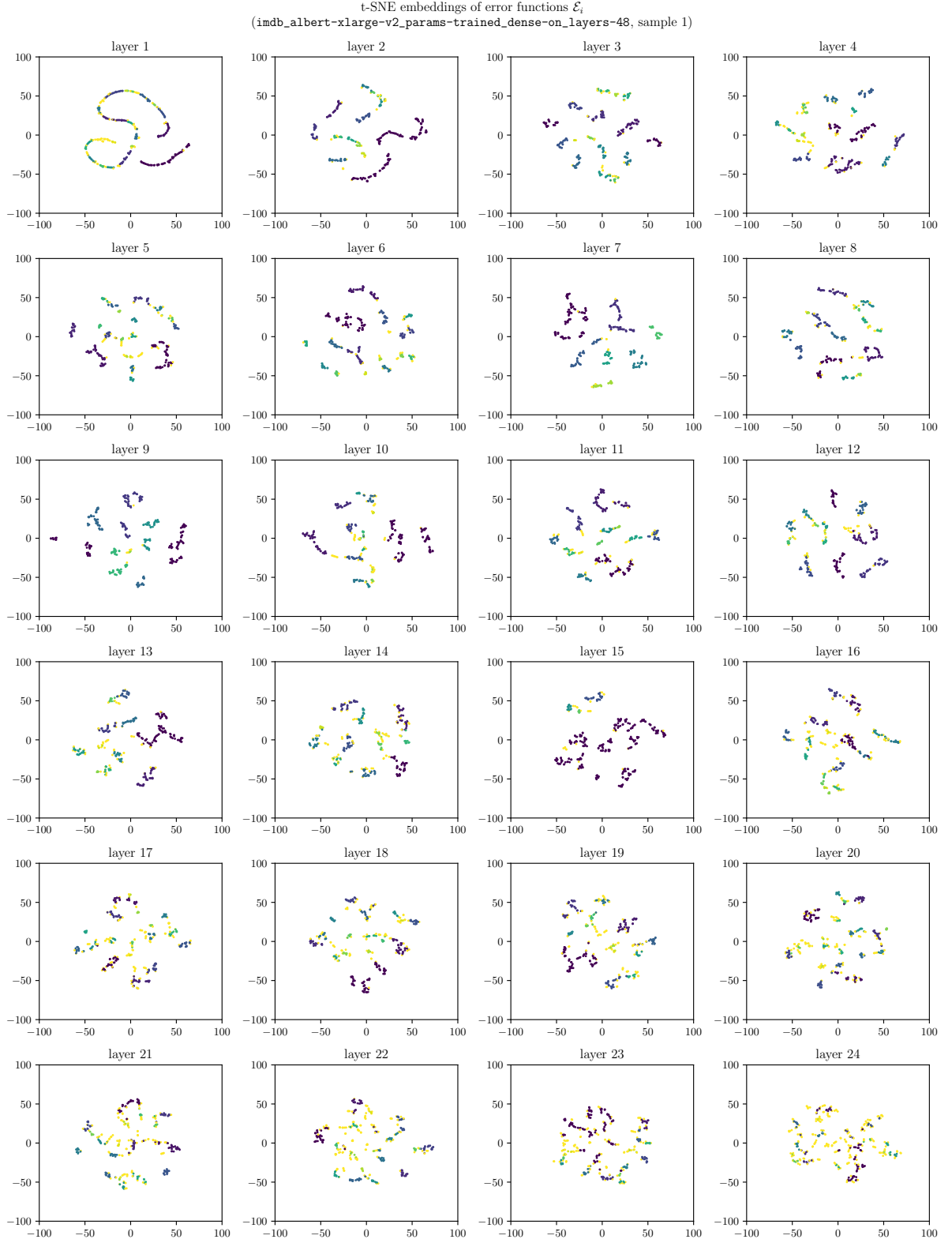


Figure 15: Error function t-SNE embeddings for layers 1 to 24. Outliers are shown in yellow.

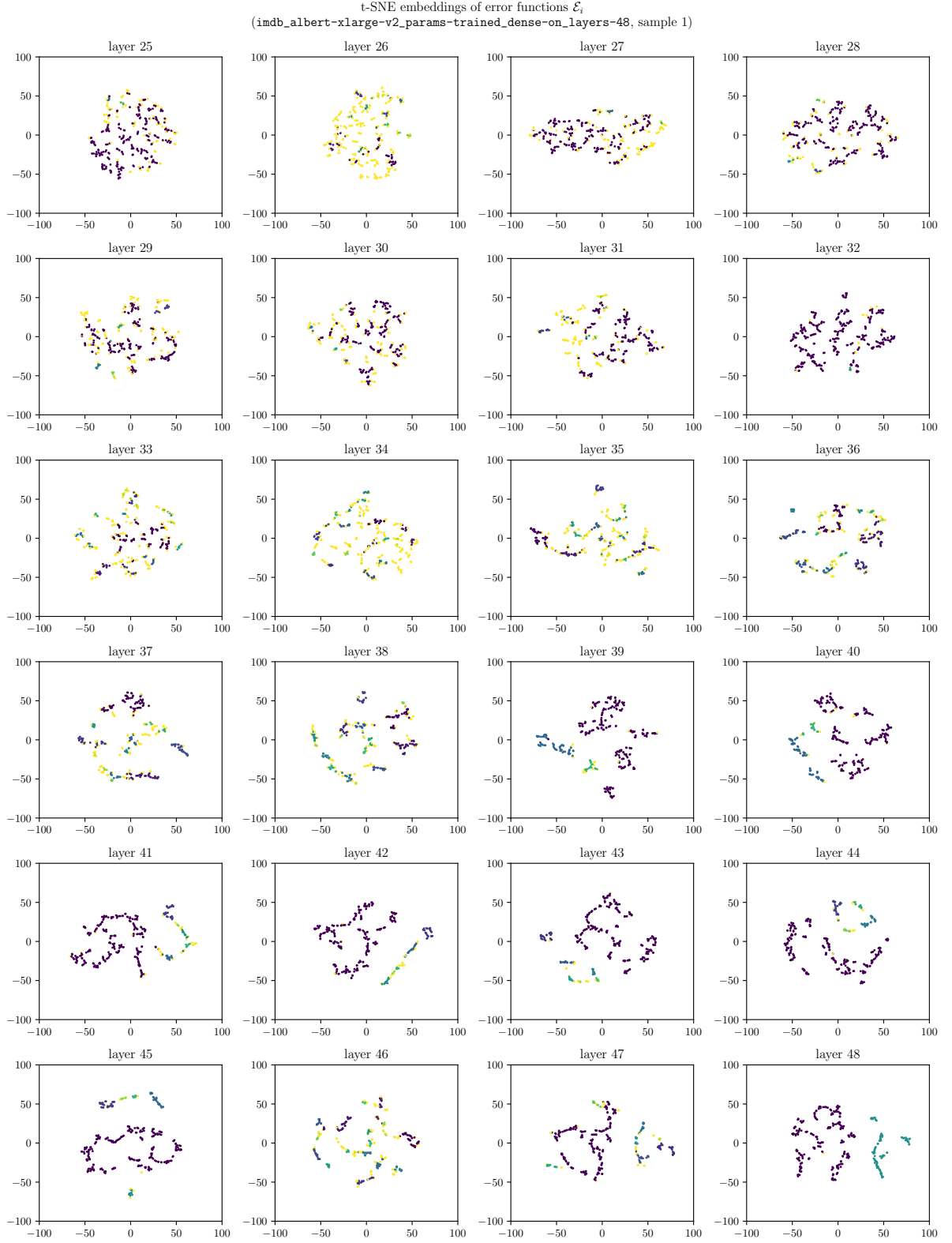


Figure 16: Error function t-SNE embeddings for layers 25 to 48. Outliers are shown in yellow.

References

- Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. Density-Based Clustering Based on Hierarchical Density Estimates. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu, editors, *Advances in Knowledge Discovery and Data Mining*, volume 7819, pages 160–172. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. Series Title: Lecture Notes in Computer Science.
- José A. Carrillo, Young-Pil Choi, Claudia Totzeck, and Oliver Tse. An analytical framework for consensus-based global optimization method. *Math. Models Methods Appl. Sci.*, 28(6): 1037–1066, 2018.
- José A. Carrillo, Nicolás García Trillos, Sixu Li, and Yuhua Zhu. FedCBO: Reaching Group Consensus in Clustered Federated Learning through Consensus-based Optimization. *Journal of Machine Learning Research*, 25(214):1–51, 2024.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Jill Burstein, Christy Doran, and Tamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD’96, pages 226–231, Portland, Oregon, August 1996. AAAI Press.
- Massimo Fornasier, Timo Klock, and Konstantin Riedl. Consensus-Based Optimization Methods Converge Globally. *SIAM J. Optim.*, 34(3):2973–3004, 2024.
- Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.
- Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on Transformers, February 2024. arXiv:2312.10794 [cs, math].
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.
- Yiping Lu, Zhuohan Li, Di He, Zhiqing Sun, Bin Dong, Tao Qin, Liwei Wang, and Tie-Yan Liu. Understanding and Improving Transformer From a Multi-Particle Dynamic System Point of View, June 2019. arXiv:1906.02762 [cs, stat].
- Davoud Moulavi, Pablo A. Jaskowiak, Ricardo J. G. B. Campello, Arthur Zimek, and Jörg Sander. Density-Based Clustering Validation. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 839–847. Society for Industrial and Applied Mathematics, April 2014.

- René Pinnau, Claudia Totzeck, Oliver Tse, and Stephan Martin. A consensus-based model for global optimization and its mean-field limit. *Math. Models Methods Appl. Sci.*, 27(1):183–204, 2017.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models, 2023. Version Number: 2.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.