Transformers meet Stochastic Blockmodels: Attention with Data-Adaptive Sparsity and Cost

Anonymous Author(s)
Affiliation
Address
email

Abstract

To overcome the quadratic cost of self-attention, recent works have proposed various sparse attention modules, most of which fall under one of two groups: 1) sparse attention under a hand-crafted patterns and 2) full attention followed by a sparse variant of softmax such as α-entmax. Unfortunately, the first group lacks adaptability to data while the second still requires quadratic cost in training. In this work, we propose SBM-Transformer, a model that resolves both problems by endowing each attention head with a mixed-membership Stochastic Blockmodel (SBM). Then, each attention head data-adaptively samples a bipartite graph, the adjacency of which is used as an attention mask for each input. During backpropagation, a straight-through estimator is used to flow gradients beyond the discrete sampling step and adjust the probabilities of sampled edges based on the predictive loss. The forward and backward cost are thus linear to the number of edges, which each attention head can also choose flexibly based on the input. By assessing the distribution of graphs, we theoretically show that SBM-Transformer is a universal approximator for arbitrary sequence-to-sequence functions in expectation. Empirical evaluations under the Long Range Arena benchmark demonstrate that our model outperforms previous efficient variants as well as the original Transformer with full attention.

1 Introduction

The Transformer [34] architecture has been the go-to method for encoding sequential data, due to its superior performance in various tasks such as machine translation [26], image classification [12], and protein language modeling [29]. Its key strength stems from the multi-head attention module, where a so-called *attention score* matrix computes how contextually important one token is to another for all possible token pairs. Each Transformer layer simultaneously pools the token representations based on the attention scores, eventually returning contextualized features without sequentially traversing through the input sequence as its recurrent neural network-based predecessors [14].

A well-known drawback of the original Transformer is its high computational cost in time and memory that increases quadratically with sequence length. This is due to the full pairwise computation of attention scores, which prohibits applying it in tasks involving long-range dependencies such as document summarization [15] or high-resolution image processing [39]. Many works have thus focused on developing more efficient alternatives by exploiting fixed or learnable attention sparsity patterns [7, 38, 18, 11], low-rank approximations [35, 36], or kernelized attention modules [17, 8].

Even though the efficient alternatives hold theoretical expressibility guarantees [37], they are far from sufficient, still failing to convince practitioners to replace the original Transformer. We believe this is mostly due to their lack of adaptability. They apply the same modifications to unanimously sparsify all the attention modules across layers, without considering the tasks at hand. Such strategy

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

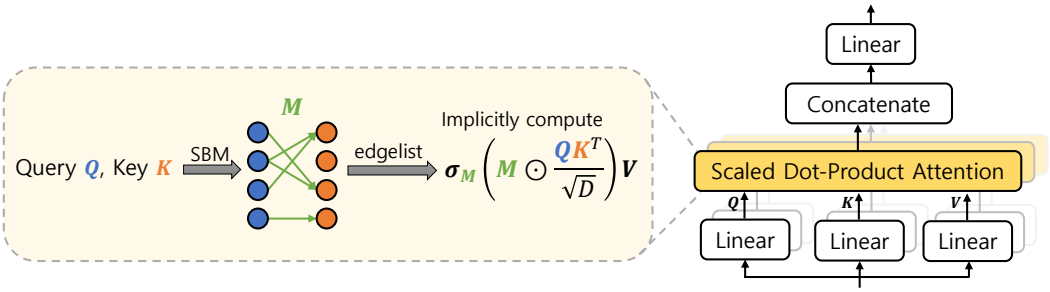

Figure 1: The attention module in SBM-Transformer. In multi-head attention, each attention head samples a bipartite graph connecting queries to keys from an underlying SBM. The adjacency of the sampled graph is used as an attention mask to compute the dot products only for the sampled edges.

imposes inductive bias too strongly and often leads to sub-optimal cost vs. performance trade-offs in downstream tasks [25]. In this work, we argue that to retain the utmost potential of the Transformers, each attention module should have the ability to flexibly choose between sparse and full attention. This is especially evident when considering many state-of-the-art systems suggest the need for a mixture of dense and sparse attention layers. For example, a qualitative analysis on pretrained BERT showed that lower layers exhibit broad dense attention while upper layers perform focused sparse attention [9]. In the case of GPT-3 [6], the Transformer blocks are manually arranged to alternate between dense and sparse attention.

To contribute to the efficient Transformers lineage, we propose SBM-Transformer, capable of adjusting its attention sparsity data-adaptively based without fully computing the attention score matrix (Figure 1). Leveraging a mixed-membership Stochastic Blockmodel (SBM) [2], each attention head samples a bipartite graph connecting queries to keys. Then, the adjacency of the sampled graph is used as an attention mask so that only attention scores corresponding to sampled edges are computed. The overall computational cost is linear in the number of edges, which can range from linear to quadratic in sequence length depending on the data and task under concern. Each attention head is equipped with its own underlying SBM, enabling the model to diversify the attention sparsity across heads and layers. By incorporating a straight-through estimator [4] in the discrete graph-sampling step, SBM-Transformer enjoys end-to-end differentiability and can find the proper attention sparsity based solely upon minimizing the predictive loss. The model can also easily be further regularized by penalizing the number of sampled edges, which results in a lighter model using less computational resources during inference. To the best of our knowledge, our method is the first Transformer architecture that can data-adaptively choose between linear to full attention with respective computational costs. To summarize, our main contributions are as follows:

- We present SBM-Transformer, a novel Transformer of which each attention head can adaptively adjust its attention sparsity as well as computational cost based on the input data.
- To demonstrate the benefit of this flexibility, we theoretically prove that SBM-Transformer retains universal approximability, and also stress-test the model under a synthetic task where full attention is required to achieve 100% accuracy.
- Extensive evaluations on the LRA benchmark show that SBM-Transformer outperforms previous efficient Transformer models as well as the vanilla Transformer with dense attention.

2 Related Work

In this section we discuss previous efficient Transformer variants and several works similar to ours with respect to adaptively learning sparse attention patterns. We also review several works on SBMs.

Efficient Transformers. Many efficient Transformers tackle to reduce the quadratic cost of multi-head attention with different approaches. While we provide only a handful of representative approaches, a much more comprehensive survey can be found in [33]. The Linear Transformer [17] achieves linear complexity by replacing the softmax with a low-rank kernelized function. Linformer [35] and Nyströmformer [36] use a similar approach by low-rank approximating the attention score matrix. Performer [8] uses positive orthogonal random features to approximate the softmax

kernel. Reformer [18] gathers similar tokens together through locality-sensitive hashing (LSH) and performs attention amongst tokens within the same bucket. Of all methods above, our method is most similar to Reformer, in the sense that we adaptively assign queries and keys into clusters and form a low-rank sparse attention pattern. However, our method performs soft-clustering with much less structural constraints, allowing each attention head to represent a wider variety of dependency structure and to adjust its sparsity towards full attention if needed.

Adaptive Sparsity. With respect to flexible training between sparse and dense attention, there exist some works that parameterize how sparse the attention pattern should be based on the input. The Adaptive Sparse Transformer [10] proposed replacing the usual softmax activation with α-entmax, in which the α parameter can be differentiably trained to adjust the activation between softmax and sparsemax activation [23]. SparseBERT [31] uses a differentiable masking technique where each attention mask is sampled from a Gumbel-sigmoid distribution using data-independent mask probability parameters. While these methods possess the flexibility to adjust between sparse and full attention based on data, they still require full computation of the attention score matrix before sparsification, and hence are unable to leverage the learned sparsity towards better model efficiency. To the best of our knowledge, ours is the first work to be able to adaptively tune its attention sparsity between sparse to full attention without requiring the explicit computation of the attention score matrix, thereby avoiding quadratic cost when possible.

Stochastic Blockmodels. The Stochastic Blockmodel (SBM) is a generative model that encodes the latent structure of graphs by grouping nodes into clusters. By modeling the cluster-membership of each node as well as inter-cluster relationships, SBMs can represent a wide variety of graph structures, which is a feature especially useful for generating new graphs or predicting missing edges in noisy data [1]. The standard SBM assigns each node to a single cluster, and the probability of an edge between two nodes strictly depends on the corresponding clusters. Several structural extensions include overlapping SBM [20] and mixed-membership SBM [2], which allow each node to be assigned to multiple clusters. The underlying SBM used by our framework mostly resembles these two variants, while the edge probability is modeled by a nonlinear function of two node embeddings rather than a bilinear one. There exist many other extensions including degree-corrected SBM [16] for multi-graphs and hierarchical SBM [27] for multiplex-graphs. Further details can be found in a recent survey [13].

3 Preliminaries: Sparse Transformers

We first introduce the full attention mechanism used in the original Transformer [34] as well as masked attention which will serve as a backbone of our approach.

3.1 Full Attention

In vanilla Transformer [34], each attention head takes a sequence of token features as input $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ where n is the sequence length and d the embedding dimension. Weight parameters $\boldsymbol{W}^Q, \boldsymbol{W}^K \in \mathbb{R}^{d \times d_h}$ and $\boldsymbol{W}^V \in \mathbb{R}^{d \times d_h}$ with head-dimension d_h first maps the input features $\boldsymbol{X}$ into query $\boldsymbol{Q}$, key $\boldsymbol{K}$, and value $\boldsymbol{V}$, respectively. Then, the *attention score matrix* is computed with scaled dot-product of queries and keys followed by row-wise softmax activation $\sigma(\cdot)$. Note that explicit computation of this matrix is the main bottleneck of full attention, incurring $\mathcal{O}(n^2)$ asymptotic cost in both time and memory. The value features $\boldsymbol{V}$ are then pooled based on the attention scores, returning the output token representations. Altogether, the operation performed by each attention head can be written as

$$\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}^Q, \;\; \boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}^K, \;\; \boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}^V \tag{1}$$

$$\texttt{Attn}(\boldsymbol{X}) = \sigma\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_h}}\right)\boldsymbol{V}. \tag{2}$$

3.2 Masked Attention

One way to remove the quadratic bottleneck from the attention score matrix is to apply a binary mask $\boldsymbol{M} \in \{0,1\}^{n \times n}$ and compute the scaled dot-products $\boldsymbol{Q}_i\boldsymbol{K}_j/\sqrt{d_h}$ only if $\boldsymbol{M}_{ij} = 1$. In presence of

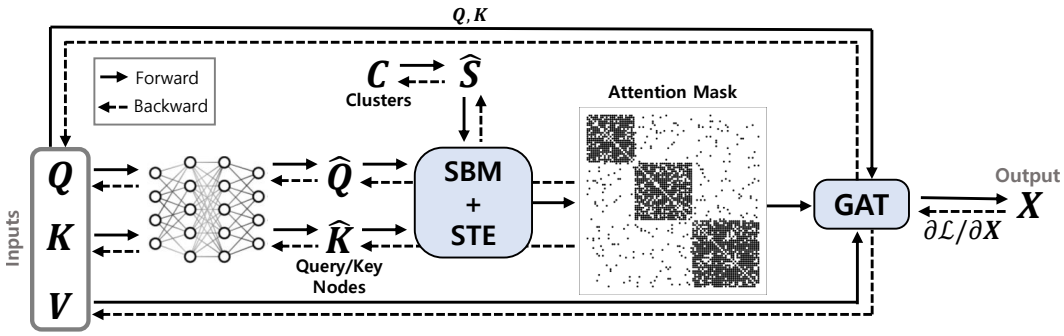

Figure 2: An illustration of the attention mechanism in SBM-Transformer. Each head first maps queries and keys to the node representation space through a shared MLP. The graph sampling module samples an attention mask from a Stochastic Blockmodel (SBM) parameterized by the node and cluster embeddings. The discrete sampling step is differentiable via a Straight-Through Estimator (STE). Given the mask, the output is computed using a Graph Attention Network (GAT) module.

an attention mask, the operation is modified to

$$\texttt{Attn}_{\text{mask}}(\boldsymbol{X}, \boldsymbol{M}) = \sigma_{\boldsymbol{M}}\left(\boldsymbol{M} \odot \frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_h}}\right)\boldsymbol{V} \tag{3}$$

$$\sigma_{\boldsymbol{M}}(\boldsymbol{A})_{ij} \coloneqq \frac{\exp(\boldsymbol{A}_{ij})}{\sum_{k,\boldsymbol{M}_{ik}=1}\exp(\boldsymbol{A}_{ik})} \tag{4}$$

where $\odot$ indicates entry-wise multiplication. Note the normalizing term in masked-softmax $\sigma_{\boldsymbol{M}}(\cdot)$ only sums up unmasked terms, ensuring that each (i, j)-th attention score survives as nonzero if and only if $\boldsymbol{M}_{ij} = 1$. Most sparsity-based efficient Transformers fall under this formulation, while using different methods to either manually fix or learn the mask $\boldsymbol{M}$. For instance, local attention [7, 3, 38] with a sliding window sets $\boldsymbol{M}_{ij} = 1$ if $|i-j| < c$ for some context window size c while Reformer [18] sets $\boldsymbol{M}_{ij} = 1$ if $\boldsymbol{Q}_i$ and $\boldsymbol{K}_j$ are hashed into the same bucket.

4 Our Method: SBM-Transformer

Here we discuss the details of SBM-Transformer (Figure 2). We first illustrate the forward step of our attention module and how the underlying SBM [2] of each head, from which we sample our attention masks, is parameterized by the input tensors. We then discuss how the model enables end-to-end differentiability despite the discrete graph sampling step.

4.1 Forward step with the Stochastic Blockmodel

In our framework, we view the attention mask $\boldsymbol{M}$ as an adjacency matrix of a bipartite graph that connects queries to keys, and let each attention head sample an adjacency matrix that best represents the contextual dependencies amongst input tokens. In order to efficiently sample adjacency matrices while avoiding the quadratic cost, the distribution of graphs must first be parameterized with a sub-quadratic number of latent variables. Stochastic blockmodels fit perfectly for our purpose as it models graphs that are low-rank structured with k latent clusters, allowing full parameterization with $\mathcal{O}(nk)$ memory. More concretely, the SBM distribution is defined by two nonnegative node-to-cluster memberships $\boldsymbol{Y}, \boldsymbol{Z} \in \mathbb{R}_+^{n\times k}$ and a so-called block matrix $\boldsymbol{B} \in \mathbb{R}_+^{k\times k}$ that stores the inter-cluster connection probabilities. The probability of node i being connected to node j is computed as $p(i, j) = \boldsymbol{Y}_i\boldsymbol{B}\boldsymbol{Z}_j^T$. Equivalently, the expectation of the adjacency matrix sampled from $\boldsymbol{A} \sim SBM(\boldsymbol{Y}, \boldsymbol{B}, \boldsymbol{Z})$ can be written as $\mathbb{E}[\boldsymbol{A}] = \boldsymbol{Y}\boldsymbol{B}\boldsymbol{Z}^T$.

For proper parameterization of the SBM, we must infer the nonnegative node-memberships and block matrix from the queries and keys. To do so, we equip each attention head a 2-layer $\text{MLP}_{d_h\to d_h}$ with ReLU activation, and a set of k trainable cluster-embeddings $\boldsymbol{C} \in \mathbb{R}^{k\times d_h}$. First, our model computes the block matrix $\hat{\boldsymbol{S}} \in \mathbb{R}_+^{k\times k}$ by taking dot products amongst cluster-embeddings $\boldsymbol{C}$ followed by a

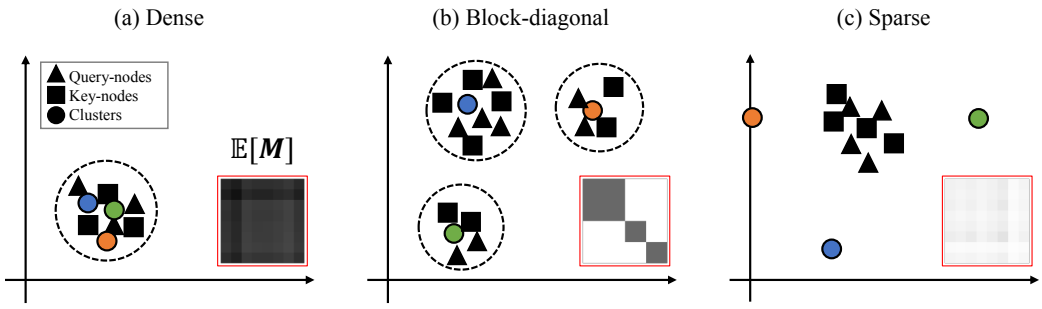

Figure 3: Representative examples from the SBM and resulting mask expectations (darker grid indicates closer to 1). (a) The expected mask is dense if all nodes and clusters are collapsed within a small region. (b) Clear-cut groups in the embedding space induce a block-diagonal mask. (c) Clusters located far apart from nodes lead to sparse masks.

2-dimensional softmax activation. The node embeddings are obtained by processing each query and key through the $\text{MLP}_{d_h \to d_h}$, mapping token representations into the node representation space. The memberships of query and key nodes, which we denote by $\hat{\boldsymbol{Q}}$ and $\hat{\boldsymbol{K}}$, are then inferred by taking dot products of node and cluster embeddings, followed by a sigmoid function. The block matrix $\hat{\boldsymbol{S}}$, query node-memberships $\hat{\boldsymbol{Q}}$, and key node-memberships $\hat{\boldsymbol{K}}$ altogether provide a well-defined parameterization for the SBM. Thus, a bipartite graph adjacency $\boldsymbol{M} \in \{0,1\}^{n \times m}$ can be sampled from $\boldsymbol{M} \sim SBM(\hat{\boldsymbol{Q}}, \hat{\boldsymbol{S}}, \hat{\boldsymbol{K}})$ with expectation $\mathbb{E}[\boldsymbol{M}] = \hat{\boldsymbol{Q}}\hat{\boldsymbol{S}}\hat{\boldsymbol{K}}^T$: the probability of connecting query $\boldsymbol{Q}_i$ to key $\boldsymbol{K}_j$ equals $p(i,j) = \hat{\boldsymbol{Q}}_i\hat{\boldsymbol{S}}\hat{\boldsymbol{K}}_j^T$. Formally, the graph sampling procedure can be written as

$$\hat{\boldsymbol{S}} = \texttt{softmax}(\boldsymbol{C}\boldsymbol{C}^T) \tag{5}$$

$$\hat{\boldsymbol{Q}} = \texttt{sigmoid}(\text{MLP}_{d_h \to d_h}(\boldsymbol{Q})\boldsymbol{C}^T) \tag{6}$$

$$\hat{\boldsymbol{K}} = \texttt{sigmoid}(\text{MLP}_{d_h \to d_h}(\boldsymbol{K})\boldsymbol{C}^T) \tag{7}$$

$$\boldsymbol{M} \sim SBM(\hat{\boldsymbol{Q}}, \hat{\boldsymbol{S}}, \hat{\boldsymbol{K}}) \tag{8}$$

For the last sampling step, we incorporate a fast random graph sampling algorithm `fastRG` (Alg. 1, [30]) that can sample graphs from a SBM in time and memory asymptotically linear in the number of edges. One advantage of `fastRG` is that each edge can be sampled in parallel, allowing high efficiency with the help of multiprocessing. A more significant feature of the method is that the number of edges, which determines the overall cost, is sampled from a Poisson distribution with input-dependent mean (Line 4). Thus, the model can dynamically adjust its computational cost between linear and quadratic in sequence length based on the data.

Figure 3 shows example placements of nodes and clusters on the d_h-dimensional space to show how the sparse structure is determined. If all nodes and clusters are gathered closely, then all entries in $\hat{\boldsymbol{Q}}$ and $\hat{\boldsymbol{K}}$ become close to 1, resulting in $p(i,j) \approx 1$ for all i, j and hence a dense $\boldsymbol{M}$. If clusters are well-separated but each surrounded by some set of nodes, $\hat{\boldsymbol{S}}$ becomes close to diagonal while each row in $\hat{\boldsymbol{Q}}$ and $\hat{\boldsymbol{K}}$ is close to a one-hot vector indicating the cluster nearby. Such setting leads to a block diagonal mask similar to LSH bucketing of Reformer [18]. Lastly, if all clusters are far apart from the nodes, both $\hat{\boldsymbol{Q}}$ and $\hat{\boldsymbol{K}}$ approximately equal zero, zeroing out all the edge probabilities.

4.2 Backward Step with Straight-Through Estimator

The graph sampling procedure is naturally a discrete operation. Thus, a naive backpropagation can not learn the proper parameterization for the SBM that minimizes the predictive loss. To cope with this non-differentiability, we incorporate a Straight-Through Estimator (STE) [4] to pass the gradient beyond the discrete sampling step. The STE enables providing the gradient $\partial\mathcal{L}/\partial\boldsymbol{M}_{ij}$ to the probability for each sampled edge (i,j) (Eqn. 9). It works as if we had used a continuous mask $\boldsymbol{M} \odot \mathbb{E}[\boldsymbol{M}]$ that stores the probability of each sampled edge instead of the binary mask $\boldsymbol{M}$ during forward propagation. This way, the probabilities of sampled edges can be learned end-to-end based on the loss. The gradients provide information on whether each sampled edge was useful or not for prediction.

Algorithm 1: `fastRG`$(\boldsymbol{Y}, \boldsymbol{B}, \boldsymbol{Z})$[30]

Input : $\boldsymbol{Y} \in \mathbb{R}_+^{n\times k}$, $\boldsymbol{B} \in \mathbb{R}_+^{k\times k}$, $\boldsymbol{Z} \in \mathbb{R}_+^{n\times k}$
Output : $\boldsymbol{M} \in \{0,1\}^{n\times n}$ with $\mathbb{E}[\boldsymbol{M}] = \boldsymbol{Y}\boldsymbol{B}\boldsymbol{Z}^T$

Compute diagonal matrices $\boldsymbol{D_Y} = (\text{diag}(\mathbf{1}\boldsymbol{Y}))^{-1}$ and $\boldsymbol{D_Z} = (\text{diag}(\mathbf{1}\boldsymbol{Z}))^{-1}$
Column-normalize $\overline{\boldsymbol{Y}} = \boldsymbol{Y}\boldsymbol{D_Y}^{-1}$ and $\overline{\boldsymbol{Z}} = \boldsymbol{Z}\boldsymbol{D_Z}^{-1}$
Compute $\overline{\boldsymbol{B}} = \boldsymbol{D_Y}\boldsymbol{B}\boldsymbol{D_Z}$
Sample number of edges $m \sim \text{Poisson}(\mathbf{1}\overline{\boldsymbol{B}}\mathbf{1}^T)$
Initialize $\boldsymbol{M} = \mathbf{0}$
for $i = 1 : m$ **do**
  Sample (U, V) from $\{1, \ldots, k\} \times \{1, \ldots, k\}$ with $Pr(U = u, V = v) \propto \overline{\boldsymbol{B}}_{uv}$
  Sample source I from $\{1, \ldots, n\}$ with $Pr(I = i) = \overline{\boldsymbol{Y}}_{iU}$.
  Sample destination J from $\{1, \ldots, n\}$ with $Pr(J = j) = \overline{\boldsymbol{Z}}_{jV}$
  Set $\boldsymbol{M}_{IJ} = 1$.
end

$$\frac{\partial \mathcal{L}}{\partial p_{ij}} := \frac{\partial \mathcal{L}}{\partial \boldsymbol{M}_{ij}} = \begin{cases} \dfrac{\partial \mathcal{L}}{\partial \boldsymbol{A}_{ij}} \cdot \dfrac{\boldsymbol{Q}_i \boldsymbol{K}_j^T}{\sqrt{d_h}} & \text{if } \boldsymbol{M}_{ij} = 1 \\ 0 & \text{otherwise} \end{cases} \quad \text{where } \boldsymbol{A} := \boldsymbol{M} \odot \frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_h}} \tag{9}$$

Random Edge Exploration. While this approach enables backpropagation in the same $\mathcal{O}(m)$ cost as in the forward step, this comes at the expense of not being able to propagate information through edges not sampled. This can be problematic when an edge probability accidentally collapses to zero, after which the edge becomes unlikely to ever be sampled even when it may be useful for the prediction task at hand. Therefore, we add a small perturbation $\delta > 0$ to each edge probability p_{ij}, allowing the model to explore new edges and resuscitate their sampling probabilities if necessary. We find that a δ as small as 0.01 significantly helps in practice, and thus use this exploration scheme during training for our experiments.

Wouldn't the model always prefer full attention? Note that the gradient $\partial\mathcal{L}/\partial p_{ij}$ can be positive, which suppresses the probability of edge (i, j). At first, it may seem counter-intuitive why the model would ever limit itself to using fewer edges during training. One explanation is that masked attention provides an easy way to reduce attention scores under finite head dimensions. Under full attention, it is known that the representational space of attention score matrices is limited by the head dimension and softmax activation [5]. This limitation inevitably introduces unwanted noise in the attention scores especially when working with long sequences. In SBM-Transformer, however, the structural sparsity in masked attention introduces another dimension that induces a larger space of row-stochastic matrices (full attention is a special case of masked attention where $\boldsymbol{M}_{ij} = 1$ for all i, j). Therefore, it is reasonable that the model may encourage sparsity to leverage the additional expressiveness assuming the loss landscape has local optima within the sparse attention regime. Our experiments on the LRA benchmark show that this is indeed the case, as our SBM-Transformer converges to an average attention sparsity of 20% to 30% while outperforming Transformer with full attention. We also show in the experiment that we can easily incorporate additional regularization that further encourages the sparse mask.

4.3 SBM-Transformer is a Universal Approximator

Following the theoretical analysis of BigBird [38], we demonstrate that the mixed-membership SBM with adaptability and stochasticity provides the same level of expressibility as full attention. For brevity, we briefly discuss the main idea of the proof, and defer the details to the appendix.

Theorem 1. *Let $f \in \mathcal{F}$ be class of continuous sequence-to-sequence functions. $\mathcal{T}_{SBM}^{h,r,m}$ denote the class of SBM-Transformers with h attention heads, m head dimension, and r dimensions in hidden layers. Then for any $\epsilon > 0$ and $1 \le p < \infty$, there exists a function $g \in \mathcal{T}_{SBM}^{h,m,r}$ such that*

$$\int_{\mathbb{D}} \|f(\boldsymbol{X}) - \mathbb{E}[g(\boldsymbol{X})]\|_p^p d\boldsymbol{X} \le \epsilon \tag{10}$$

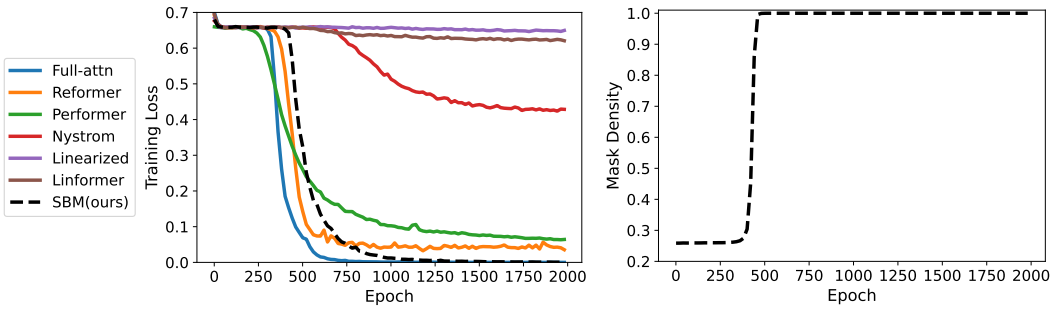

Figure 4: Loss curves (left) and overall sparsity of SBM-Transformer (right) during training for the synthetic task. SBM-Transformer successfully converges to zero loss by tuning itself towards full attention.

Theorem 1 from [38] states that sparse attention enjoys expressibility if the star graph with one global relay token is contained within the sparsity graph. In our setting, the star graph can be viewed as a rank-2 graph structure where the global relay token is in its own cluster, and all other tokens belong to another. As long as each attention head uses more than 2 clusters, there exists an SBM which can sample the star graph in expectation, hence our model achieves universal approximability.

5 Experiments

For empirical evaluations, we first use a synthetic task to show that our model is flexible enough to learn towards full attention when needed in contrast to previous works. We then experiment on Long Range Arena (LRA) [32], a benchmark widely used to assess the capacity of efficient Transformers in learning long-range contexts across different modalities. All experiments were run on a remote GCP server equipped with 16 NVIDIA A100 Tensor Core GPUs.

5.1 Synthetic Task: Finding Repeated Tokens

Dataset. We formulate a token-level binary classification task as follows: each input sequence consists of N integers, each of which is uniformly sampled from $\{1, 2, \ldots, N\}$. We use $N = 256$ in our setup. The prediction target is a sequence of equal length, where each token is labeled 1 if there exists a duplicate somewhere within the sequence, and 0 otherwise. Below is a simple example with $N = 8$ that illustrates the task. We measure the performance of models via binary cross-entropy loss.

$$\text{Input: } \texttt{1 4 3 7 3 2 3 1} \Rightarrow \text{Target: } \texttt{1 0 1 0 1 0 1 1}$$

Methods. For this task, we compare SBM-Transformer with $k = 128$ clusters against various efficient Transformers: Linear Transformer [17], Linformer [35], Reformer [18], Performer [8], and Nyströmformer [36]. Across all methods, we use a single-layer and single-head architecture with 32 hidden dimensions. Note that due to this constrained setting, the sole head must perform full attention to compare each token to all the others and attain 100% accuracy. All models are trained for 2000 epochs where a new batch of sequences is sampled for every epoch. We use a batch size of 256 and learning rate of 1e-3.

Results. Figure 4 shows the training loss curves of each baseline method as well as SBM-Transformer. Full-attention quickly converges to 100% accuracy as expected since it computes all possible pairwise interactions by default. Other models that apply low-rank or kernelized attention are unable to achieve the same level of accuracy, due to limited expressibility under the constrained setting. Though SBM-Transformer converges more slowly compared to full-attention, it demonstrates the ability to drive itself towards full-attention, eventually attaining zero loss.

5.2 Long Range Arena (LRA)

To demonstrate that the flexible inductive bias of SBM-Transformer is effective in modeling long-range dependencies, we test SBM-Transformer against previous work on the LRA benchmark. We

Model	LISTOPS(2K)	TEXT(3K)	RETRIEVAL(4K)	IMAGE(1K)	PATHFINDER(1K)	Avg.
Full-attention [34]	37.22	64.93	79.55	40.38	74.26	59.27
Linearized [17]	37.46	64.90	**81.10**	38.48	74.61	59.31
Reformer [18]	22.92	64.70	77.25	**43.65**	70.28	55.76
Performer [8]	18.25	65.00	79.01	39.80	70.79	54.57
Linformer [35]	**38.44**	56.28	78.09	39.53	67.62	55.99
Nyströmformer [36]	37.22	<u>65.46</u>	79.35	<u>43.07</u>	71.97	<u>59.41</u>
SBM-Transformer (ours)	37.45 (20.09%)	**65.79 (26.10%)**	<u>80.00 (29.46%)</u>	41.31 (20.49%)	**75.12 (18.56%)**	**59.93**

Table 1: LRA benchmark results. The input sequence lengths are shown next to each task. For SBM-Transformer, we report the the average attention sparsity across all layers and heads during test time in parentheses. Bold and underlined results indicate best and 2nd best test accuracy for each task, respectively.

also test how the performance is affected with respect to changing number of clusters of the SBMs and applying a sparsity-based regularizer.

Dataset. LRA [32] consists of five different testbeds with varying modalities: LISTOPS [24] is a 10-way classification task to map a sequence to its corresponding solution. Each sequence consists of single-digit numbers and 4 different set operations. TEXT [22] is a binary classification task where byte-level IMDB movie reviews must be classified into one of two sentiments: positive or negative. RETRIEVAL [28] is also a char-level binary classification task. Unlike TEXT, two sequences from ACL Anthology papers are given as input, and the model must predict whether there exists a citation link between them. IMAGE [19] is a 10-way classification task mapping flattened pixel-sequences from CIFAR-10 to its class. PATHFINDER [21] provides flattened pixel-sequences from an image and the model must decide whether two circles are connected by a dashed line. For this benchmark, we use the PyTorch implementation of LRA provided by the authors of Nyströmformer [36] and adhere to the same train-test splits. Performance in all five tasks is measured using classification accuracy.

Methods. We compare SBM-Transformer against the same baseline methods as with the synthetic task above. For fair comparison, we set all Transformer models to use the default setting suggested in previous work [36], which fixes 2 layers, 2 attention heads, and 64 embedding dimensions. The output token representations are mean-pooled to obtain the sequence representation for all tasks. For SBM-Transformer, we use $k = 128$ clusters in all attention heads. More details on the architecture setups can be found in the appendix.

Results. Table 1 shows the test accuracies of each method. Our SBM-Transformer achieves the best overall performance, ranking first in two tasks, and second in one other. SBM-Transformer also outperforms full attention in all five tasks while computing 30% or less attention scores on average, which supports our claim that masked attention with partial attention score computations can be preferred over full attention depending on the task. With respect to the attention mask structure, we find that flexibility of SBM is indeed beneficial, as Reformer struggles in LISTOPS, most likely due to the inability of block-diagonal masks to model hierarchical contexts.

Mask Density Regularization. To test if the model can effectively optimize the tradeoff between computational cost and performance, we also test the model with the same hyperparameter setting under a sparsity-based regularizer that discourages excessive use of query-key edges. We penalize each sampled edge by adding to the predictive loss a weighted regularization term $\lambda\mathcal{L}_s$, where $\mathcal{L}_s$ denotes the average density of the sampled mask across all attention heads in the model. Table 2 shows the performance of SBM-Transformer across varying regularization weights λ. Surprisingly, the model retains competitive performance under the constrained setting compared to the model trained without any regularization, while reducing the average mask density significantly. The model even outperforms the unregularized model when trained under a small regularization weight $\lambda = 10^{-4}$, indicating that the signal from $\lambda\mathcal{L}_s$ does not severely interfere with that from prediction loss. It may also be the case similar local optima are shared across sparse attention regions with varying density in the loss landscape, and the regularization term is able to drive the model towards finding optimal attention scores with smaller density.

Layerwise Diversity in Sparsity. We also compare the densities of masks sampled at each layer of SBM-Transformer during test time to examine whether our model is capable of diversifying sparsity across layers for better performance. Recall that this allows models to gather information in different levels, as seen in pretrained BERT where lower layers focus on the overall content via dense attention

λ	LISTOPS(2K)	TEXT(3K)	RETRIEVAL(4K)	IMAGE(1K)	PATHFINDER(1K)	Avg.
0	37.45 (20.09%)	**65.79 (26.10%)**	80.00 (29.46%)	41.31 (20.49%)	75.12 (18.56%)	59.93
10^{-4}	37.76 (10.48%)	65.48 (26.26%)	79.93 (24.62%)	41.35 (10.70%)	**75.46 (5.16%)**	**60.00**
10^{-3}	**38.23 (10.46%)**	65.18 (26.03%)	80.00 (21.70%)	41.17 (24.60%)	74.49 (3.82%)	59.81
10^{-2}	38.20 (2.95%)	65.59 (22.43%)	**80.44 (6.99%)**	**42.20 (3.95%)**	72.79 (3.76%)	59.84
10^{-1}	37.76 (1.15%)	64.48 (10.62%)	79.46 (2.49%)	41.35 (1.33%)	73.79 (2.61%)	59.37

Table 2: LRA results of SBM-Transformer with increasing sparsity regularization weight λ. Bold results indicate best accuracy for each task. Applying a small sparsity regularizer helps in boosting overall sparsity as well as performance.

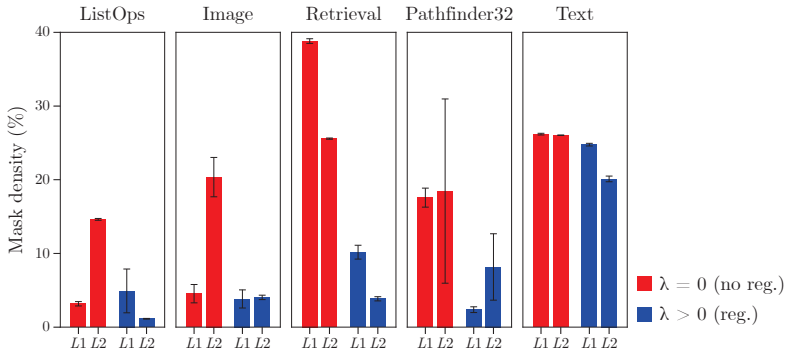

Figure 5: Average and standard deviation of density of masks sampled across the test set for each LRA task. The x-axis indicates the lower (L1) and upper (L2) layers and each bar represents the density averaged between the two attention heads in each layer.

while upper layers gather syntactic information with tree-like patterns [9]. For each of the five tasks, we pick two highest-performing models (one for unregularized and another for regularized) for measurement. Figure 5 shows the average layer-wise mask densities of unregularized and regularized SBM-Transformers across different tasks. We find that under no regularization, the two layers can differ by more than 10% in tasks such as LISTOPS and IMAGE. This may be due to the hierarchical and compositional structure of the two tasks. We also find that the variation is relatively low in TEXT with densities around 25%, indicating that the task requires broad attention overall. Lastly, the standard deviation is extremely large in upper layers for PATHFINDER, showing that it samples a wide variety of masks depending on the input.

6 Conclusion

We propose SBM-Transformer, an efficient Transformer that can data-adaptively choose its attention sparsity between sparse and full attention without the need to explicitly compute the full attention score matrix beforehand. Theoretically, we show that our model enjoys the same expressibility as the original Transformer due to the flexibility and stochasticity of the latent SBM. Empirical experiments on LRA show that our model outperforms previous state-of-the-art efficient Transformers.

Nonetheless, there are limitations from an engineering perspective as operations with sparse tensors are not GPU-friendly. It often results in longer runtimes compared to dense counterparts even though the memory usage is lower. Many sparsity-based attention mechanisms have instead used block-sparse attention that is much more amenable for GPU computation [38, 7, 3]. We still believe this work is valuable as it enables multi-head attention with long sequences, and practitioners can easily impose constraints on the number of edges to sample, based on the computational resource available in their environment. We hope to see more GPU-friendly tensor operations optimized for fine-grained sparsity in the future, at which point the value of this work will increase even further.

For future work, we hope to extend this work further towards a higher-level Neural Architecture Search (NAS) in the context of Transformers where the model not only flexibly tunes the attention sparsity, but also adjusts the number of layers and attention heads during training to autonomously optimize the tradeoff between model size and predictive performance. As we propose a foundational replacement for the scaled dot-product attention module in the Transformer architecture, we do not expect any immediate negative societal impact due to this work.

References

[1] E. Abbe. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017.

[2] E. M. Airoldi, D. Blei, S. Fienberg, and E. Xing. Mixed membership stochastic blockmodels. *Advances in neural information processing systems*, 21, 2008.

[3] I. Beltagy, M. E. Peters, and A. Cohan. Longformer: The long-document transformer. *CoRR*, abs/2004.05150, 2020.

[4] Y. Bengio, N. Léonard, and A. C. Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *CoRR*, abs/1308.3432, 2013.

[5] S. Bhojanapalli, C. Yun, A. S. Rawat, S. J. Reddi, and S. Kumar. Low-rank bottleneck in multi-head attention models. *CoRR*, abs/2002.07028, 2020.

[6] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020.

[7] R. Child, S. Gray, A. Radford, and I. Sutskever. Generating long sequences with sparse transformers. *CoRR*, abs/1904.10509, 2019.

[8] K. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlós, P. Hawkins, J. Davis, A. Mohiuddin, L. Kaiser, D. Belanger, L. J. Colwell, and A. Weller. Rethinking attention with performers. *CoRR*, abs/2009.14794, 2020.

[9] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning. What does BERT look at? an analysis of bert's attention. *CoRR*, abs/1906.04341, 2019.

[10] G. M. Correia, V. Niculae, and A. F. T. Martins. Adaptively sparse transformers. *CoRR*, abs/1909.00015, 2019.

[11] G. Daras, N. Kitaev, A. Odena, and A. G. Dimakis. SMYRF: efficient attention using asymmetric clustering. *CoRR*, abs/2010.05315, 2020.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[13] T. Funke and T. Becker. Stochastic block models: A comparison of variants and inference methods. *PLOS ONE*, 14(4):1–40, 04 2019.

[14] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[15] L. Huang, S. Cao, N. Parulian, H. Ji, and L. Wang. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436, 2021.

[16] B. Karrer and M. E. Newman. Stochastic blockmodels and community structure in networks. *Physical review E*, 83(1):016107, 2011.

[17] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. *CoRR*, abs/2006.16236, 2020.

[18] N. Kitaev, L. Kaiser, and A. Levskaya. Reformer: The efficient transformer. *CoRR*, abs/2001.04451, 2020.

[19] A. Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[20] P. Latouche, E. Birmelé, and C. Ambroise. Overlapping stochastic block models with application to the french political blogosphere. *The Annals of Applied Statistics*, 5(1):309–336, 2011.

[21] D. Linsley, J. Kim, V. Veerabadran, and T. Serre. Learning long-range spatial dependencies with horizontal gated-recurrent units. *CoRR*, abs/1805.08315, 2018.

[22] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[23] A. F. T. Martins, M. V. Treviso, A. Farinhas, V. Niculae, M. A. T. Figueiredo, and P. M. Q. Aguiar. Sparse and continuous attention mechanisms. *CoRR*, abs/2006.07214, 2020.

[24] N. Nangia and S. R. Bowman. Listops: A diagnostic dataset for latent tree learning. *CoRR*, abs/1804.06028, 2018.

[25] S. Narang, H. W. Chung, Y. Tay, W. Fedus, T. Févry, M. Matena, K. Malkan, N. Fiedel, N. Shazeer, Z. Lan, Y. Zhou, W. Li, N. Ding, J. Marcus, A. Roberts, and C. Raffel. Do transformer modifications transfer across implementations and applications? *CoRR*, abs/2102.11972, 2021.

[26] M. Ott, S. Edunov, D. Grangier, and M. Auli. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 1–9, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics.

[27] T. P. Peixoto. Hierarchical block structures and high-resolution model selection in large networks. *Physical Review X*, 4(1):011047, 2014.

[28] D. R. Radev, P. Muthukrishnan, and V. Qazvinian. The ACL Anthology network. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries (NLPIR4DL)*, pages 54–61, Suntec City, Singapore, Aug. 2009. Association for Computational Linguistics.

[29] R. M. Rao, J. Liu, R. Verkuil, J. Meier, J. Canny, P. Abbeel, T. Sercu, and A. Rives. Msa transformer. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8844–8856. PMLR, 18–24 Jul 2021.

[30] K. Rohe, J. Tao, X. Han, and N. Binkiewicz. A note on quickly sampling a sparse matrix with low rank expectation. *The Journal of Machine Learning Research*, 19(1):3040–3052, 2018.

[31] H. Shi, J. Gao, X. Ren, H. Xu, X. Liang, Z. Li, and J. T. Kwok. Sparsebert: Rethinking the importance analysis in self-attention. *CoRR*, abs/2102.12871, 2021.

[32] Y. Tay, M. Dehghani, S. Abnar, Y. Shen, D. Bahri, P. Pham, J. Rao, L. Yang, S. Ruder, and D. Metzler. Long range arena: A benchmark for efficient transformers. *CoRR*, abs/2011.04006, 2020.

[33] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler. Efficient transformers: A survey. *CoRR*, abs/2009.06732, 2020.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[35] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma. Linformer: Self-attention with linear complexity. *CoRR*, abs/2006.04768, 2020.

[36] Y. Xiong, Z. Zeng, R. Chakraborty, M. Tan, G. Fung, Y. Li, and V. Singh. Nyströmformer: A nyström-based algorithm for approximating self-attention. *CoRR*, abs/2102.03902, 2021.

[37] C. Yun, Y. Chang, S. Bhojanapalli, A. S. Rawat, S. J. Reddi, and S. Kumar. O(n) connections are expressive enough: Universal approximability of sparse transformers. *CoRR*, abs/2006.04862, 2020.

[38] M. Zaheer, G. Guruganesh, A. Dubey, J. Ainslie, C. Alberti, S. Ontañón, P. Pham, A. Ravula, Q. Wang, L. Yang, and A. Ahmed. Big bird: Transformers for longer sequences. *CoRR*, abs/2007.14062, 2020.

[39] P. Zhang, X. Dai, J. Yang, B. Xiao, L. Yuan, L. Zhang, and J. Gao. Multi-scale vision longformer: A new vision transformer for high-resolution image encoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2998–3008, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 6.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] See supplementary.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Table 2 and Figure 5 in Section 5.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]