A MODELS, DATASETS, AND IMPLEMENTATIONS

We present the details of our experiments and measurements.

A.1 PRETRAINED MODELS

Except for NanoGPT and models in Section E, we download and use pretrained models from Huggingface.

- GPT-2 (Radford et al., 2019): 12-layer, 12-head, 768-dim, 124M parameters, autoregressive, absolute positional encoding at 0th-layer, pretrained on OpenWebText;
- BERT (Devlin et al., 2018): 12-layer, 12-head, 768-dim, 124M parameters, masked prediction, absolute positional encoding at 0th-layer, pretrained on BooksCorpus and English Wikipedia;
- BLOOM (Scao et al., 2022): 24-layer, 16-head, 1024-dim, 560M parameters, ALiBI positional encodings (Press et al., 2021) at each layer, pretrained on 45 natural languages and 12 programming languages;
- Llama2-7B (Touvron et al., 2023): 32-layer, 32-head, 4096-dim, 7B parameters, autoregressive, Rotary positional embedding (Su et al., 2021) at every layer, pretrained on a variety of data.

Note that (i) the training objective for pretraining BERT is different from the other models, and (ii) Llama 2 uses rotary positional encoding for each layer and BLOOM uses ALiBI positional encoding—which is different from absolute positional encoding that is added at the 0-th layer (Vaswani et al., 2017).

A.2 TRAINING SMALL TRANSFORMERS

We train a few smaller transformers in this paper. Models are based on the GPT-2 architecture with adjusted parameters, and we adopt the implementation of the GitHub Project by Andrej Karpathy. The hardware we use is mainly RTX3090ti. The following experiments take 2 hours, 3 hours, and 3 hours to train respectively.

- **NanoGPT in Table 1 and 2**: The model is a Transformer with 6 layers, 6 heads, 384 dimensional embeddings, residual/embedding/attention dropout set to 0.1, weight decay set to 0.1, and a context window of 128. The dataset is Shakespeare with character-level tokenization. We train 100K iterations using the AdamW optimizer, with a batch size of 64 and a cosine scheduler (1000 step warmup) up to a learning rate of 5e-5;
- **Randomization**: Similarly, we use a Transformer with 8 layers, 8 heads, 512 dimensional embeddings, residual/embedding/attention dropout set to 0.1, weight decay set to 0.1, and a context window of 256. We train the model on the first 10K samples of OpenWebText dataset, which is tokenized using the same tokenizer as in GPT2. We train 100K iterations using the AdamW optimizer, with a batch size 64 and a cosine scheduler (1000 step warmup) up to a learning rate of 5e-5;
- **Addition**: Similarly, we use a Transformer with 8 layers, 8 heads, 512 dimensional embeddings, residual/embedding/attention dropout set to 0.1m and weight decay set to 0.1. The context window is set as the length of the longest sequence, i.e., 32 for the 10-digit addition task here. We train 100K iterations using the AdamW optimizer, with a batch size 64 and a cosine scheduler (1000 step warmup) up to a learning rate of 5e-5.

A.3 REMOVING ARTIFACTS

There are two likely artifacts in the measurements and visualization that we removed in the paper.

1. First token in a sequence. We find that a large proportion of attention is focused on the first token, which usually distorts visualization significantly. It has been known that the first token functions as a "null token", which is removed in analysis (Vig & Belinkov, 2019). We also adopt removing the first token in our measurements and visualization.

Table 3: ScreeNOT Rank Estimate for models, datasets and at each layer.

		Layer 0	Layer 1	Layer 2	Layer 3	Layer 4	Layer 5	Layer 6	Layer 7	Layer 8	Layer 9	Layer 10	Layer 11	Layer 12
	GitHub	15	16	16	16	14	11	11	9	10	10	11	11	12
BERT	**OpenWebText**	15	16	18	16	11	11	9	9	11	11	11	11	13
	WikiText	15	16	18	16	12	11	9	9	11	11	11	12	12
	GitHub	8	9	9	8	9	10	10	11	10	10	10	10	10
BLOOM	**OpenWebText**	6	10	10	11	11	10	11	11	11	11	10	10	11
	WikiText	6	8	9	10	10	11	11	11	11	11	11	10	11
	GitHub	15	14	13	12	12	11	11	10	10	10	11	11	10
GPT2	**OpenWebText**	15	13	14	12	13	11	10	10	10	10	9	9	12
	WikiText	15	14	14	12	11	11	11	11	11	11	9	10	12
	GitHub	6	10	9	8	10	8	8	9	9	9	9	8	10
Llama2	**OpenWebText**	7	10	10	11	11	10	9	10	9	8	9	8	10
	WikiText	8	10	10	10	9	8	8	8	8	8	8	8	10

Table 4: Stable rank for models, datasets and at each layer.

		Layer 0	Layer 1	Layer 2	Layer 3	Layer 4	Layer 5	Layer 6	Layer 7	Layer 8	Layer 9	Layer 10	Layer 11	Layer 12
	GitHub	9.19	7.79	5.26	4.73	4.34	3.84	3.48	3.20	2.70	2.45	2.04	1.84	1.91
BERT	**OpenWebText**	9.19	7.63	5.25	4.73	4.10	3.53	3.16	2.84	2.46	2.30	2.18	2.22	2.15
	WikiText	9.19	7.78	5.03	4.58	3.99	3.48	3.14	2.82	2.42	2.27	2.13	2.16	2.12
	GitHub	8.39	1.25	1.20	1.21	1.21	1.23	1.29	1.29	1.28	1.25	1.21	1.02	1.00
BLOOM	**OpenWebText**	8.33	1.27	1.30	1.24	1.24	1.27	1.32	1.34	1.33	1.26	1.16	1.01	1.00
	WikiText	8.42	1.27	1.28	1.30	1.31	1.34	1.41	1.43	1.41	1.32	1.22	1.01	1.00
	GitHub	2.05	1.92	1.91	1.89	1.90	1.90	1.92	1.94	1.98	2.03	2.05	1.70	1.11
GPT2	**OpenWebText**	2.05	1.92	1.91	1.89	1.88	1.88	1.88	1.90	1.91	1.96	2.02	2.24	1.49
	WikiText	2.05	1.92	1.91	1.89	1.88	1.88	1.88	1.90	1.91	1.97	2.03	2.19	1.56
	GitHub	24.87	1.00	1.00	1.00	1.00	1.00	1.00	1.01	1.01	1.01	1.02	1.03	1.17
Llama2	**OpenWebText**	52.23	1.00	1.00	1.00	1.00	1.00	1.01	1.01	1.02	1.02	1.03	1.05	1.44
	WikiText	24.70	1.00	1.00	1.01	1.01	1.02	1.03	1.05	1.09	1.16	1.20	1.26	1.30

2. Final-layer embeddings. We find that the embeddings of the final layer typically do not have a significant positional basis component. It is likely that positional information is no longer needed since last-layer embeddings are directly connected to the loss function.

A.4 Positional basis calculation

We calculate positional bases based on sampled sequences of length T from a subset of the corpus, which includes OpenWebText, WikiText, and GitHub. The implementation and weights of the pretrained models are obtained from HuggingFace.

For the curated corpus subset, we utilize the streaming version of the HuggingFace datasets and extract the first 10K samples from the train split. Then we tokenize the dataset using the same tokenizer employed by the pretrained model. The size of the final datasets vary across tasks and datasets, and we ensure that there are at least 1M tokens in each case to prevent the occurrence of overlapping sequences.

We set the context window $T = 512$ for BERT, BLOOM, and GPT-2, as this maintains the maximum context window utilized during pretraining. For Llama2, we set $T = 512$ instead of the maximum context sequence due to computational resource limitations.

B Additional empirical results for Section 2

B.1 PCA visualization

See Figure 7—Figure 11. Note: BERT displays a more complex circular shape, likely because its training objective is different from the others.

B.2 Low rank measurements

Rank estimate. We report the rank estimate for all pretrained models and datasets in Table 3. Additionally, we include the Stable rank estimate in Table 4.

Relative norm. We report the relative norm for all pretrained models and datasets in Table 5.

Spectral analysis. Recall that in Figure 2 (left), we showed the singular values plot for $\boldsymbol{P} = [\mathbf{pos}_1, \ldots, \mathbf{pos}_T]$, $\boldsymbol{Cvec} = [\mathbf{cvec}_{1,1}, \ldots, \mathbf{cvec}_{c,T}]$, $\boldsymbol{R} = [\mathbf{resid}_{1,1}, \ldots, \mathbf{resid}_{c,T}]$. Note that $\boldsymbol{P}$

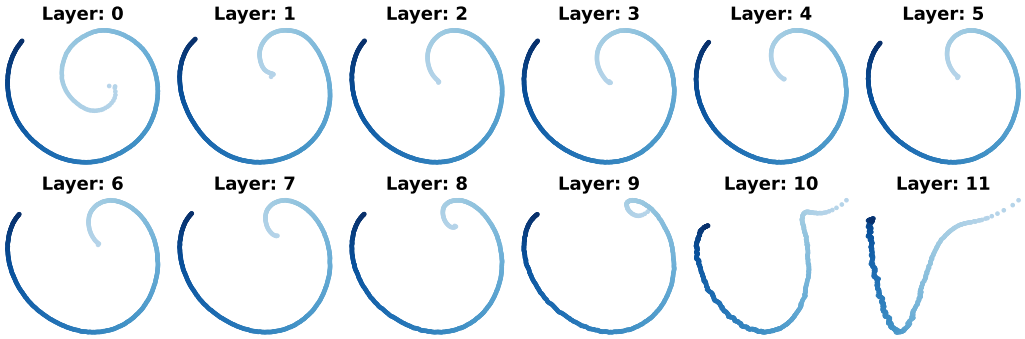

Figure 7: Top-2 principal components of positional basis; GitHub, GPT2

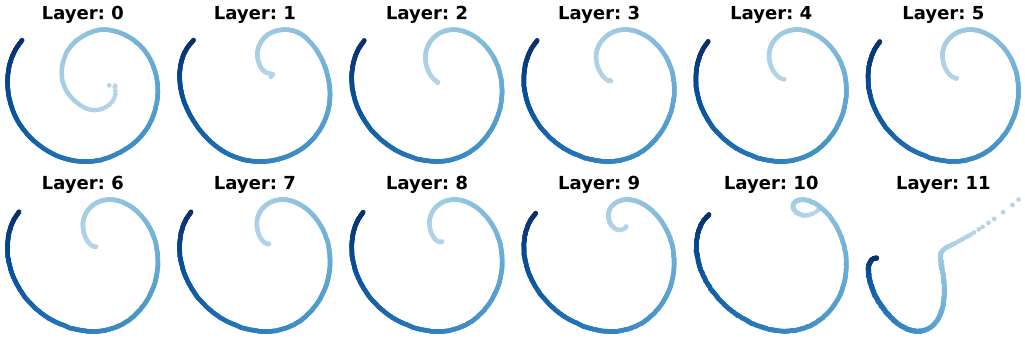

Figure 8: Top-2 principal components of positional basis; WikiText, GPT2

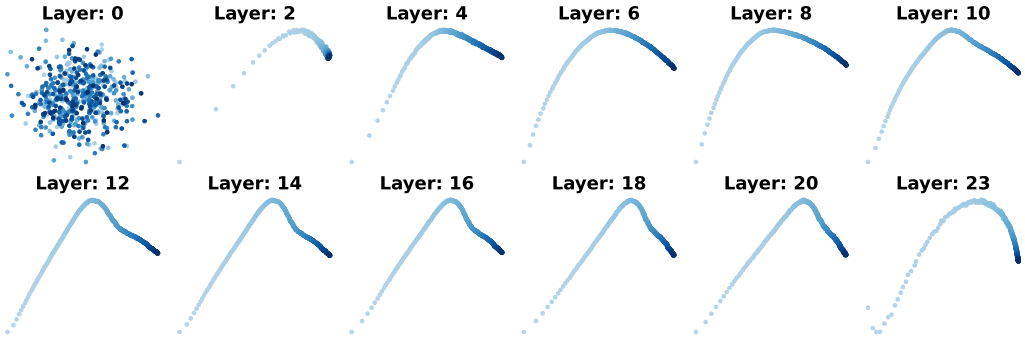

Figure 9: Top-2 principal components of positional basis; OpenWebText, BLOOM

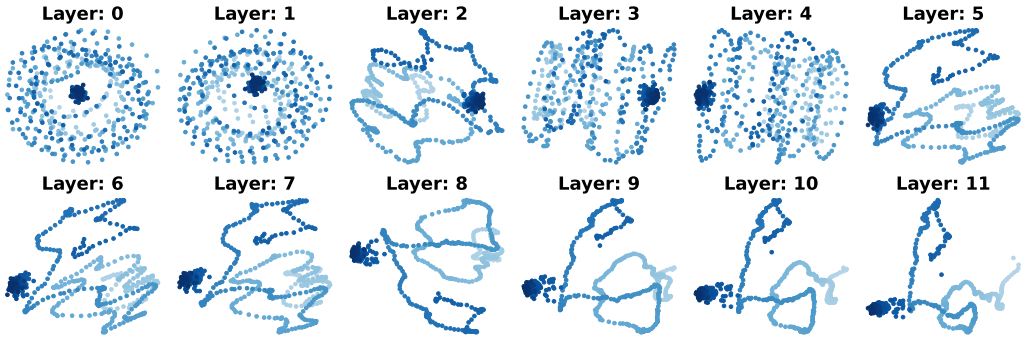

Figure 10: Top-2 principal components of positional basis; OpenWebText, BERT

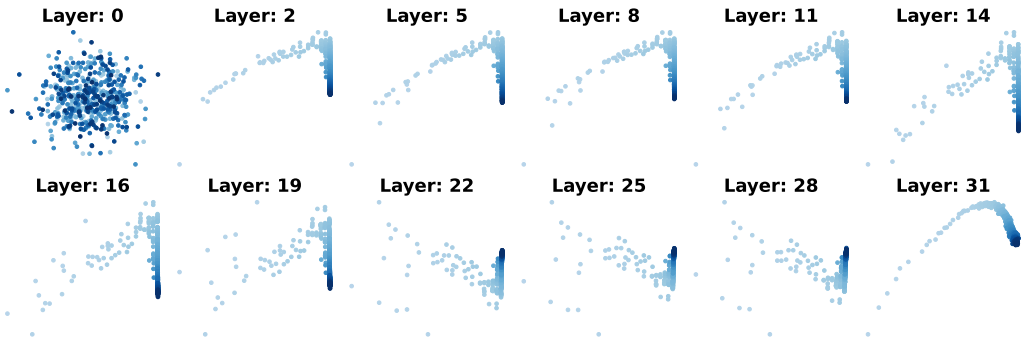

Figure 11: Top-2 principal components of positional basis; OpenWebText, Llama2

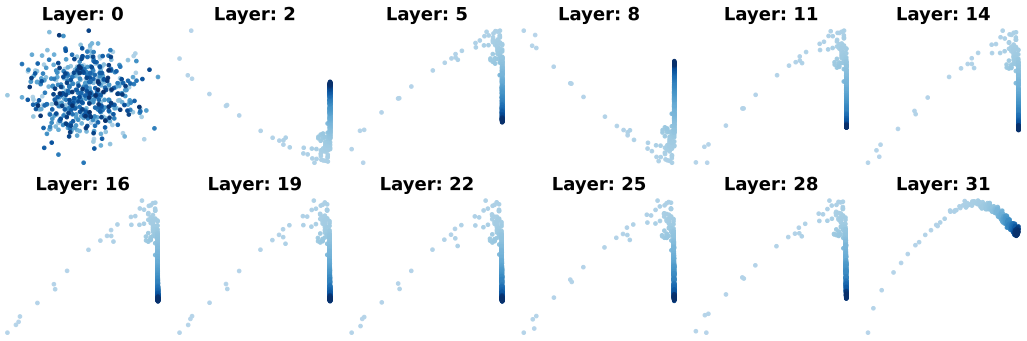

Figure 12: Top-2 principal components of positional basis; GitHub, Llama2

Table 5: Relative norm for models, datasets and at each layer.

		Layer 0	Layer 1	Layer 2	Layer 3	Layer 4	Layer 5	Layer 6	Layer 7	Layer 8	Layer 9	Layer 10	Layer 11	Layer 12
	GitHub	0.325	0.267	0.246	0.242	0.234	0.229	0.232	0.248	0.248	0.215	0.211	0.173	0.162
BERT	**OpenWebText**	0.361	0.302	0.265	0.265	0.263	0.249	0.245	0.266	0.273	0.215	0.170	0.117	0.122
	WikiText	0.356	0.293	0.267	0.265	0.263	0.250	0.246	0.269	0.278	0.217	0.172	0.118	0.122
	GitHub	0.009	0.079	0.123	0.147	0.170	0.181	0.173	0.158	0.145	0.138	0.137	0.182	0.140
BLOOM	**OpenWebText**	0.008	0.092	0.108	0.143	0.164	0.173	0.158	0.143	0.131	0.135	0.153	0.278	0.385
	WikiText	0.009	0.098	0.123	0.152	0.166	0.174	0.159	0.143	0.133	0.134	0.147	0.250	0.358
	GitHub	0.758	0.447	0.396	0.359	0.338	0.316	0.287	0.266	0.237	0.207	0.172	0.140	0.100
GPT2	**OpenWebText**	0.807	0.463	0.411	0.371	0.344	0.324	0.301	0.271	0.232	0.196	0.150	0.095	0.030
	WikiText	0.815	0.470	0.416	0.375	0.345	0.325	0.299	0.270	0.229	0.197	0.152	0.098	0.030
	GitHub	0.025	0.221	0.221	0.221	0.222	0.222	0.222	0.223	0.224	0.225	0.226	0.226	0.158
Llama2	**OpenWebText**	0.031	0.146	0.146	0.146	0.146	0.147	0.147	0.148	0.150	0.152	0.153	0.154	0.118
	WikiText	0.035	0.067	0.067	0.067	0.067	0.068	0.068	0.069	0.071	0.073	0.077	0.080	0.146

has T columns while $\boldsymbol{Cvec}$ and $\boldsymbol{R}$ has CT columns. In Figure 2 (left), we downsampled $\boldsymbol{Cvec}$ and $\boldsymbol{R}$ to match the number of columns of $\boldsymbol{P}$. Alternatively, we also tried multiplied $\boldsymbol{P}$ by $\sqrt{C}$ and got similar results.

B.3 Fourier analysis

See Figure 13—Figure 18. Compared with Figure 2 (right), for completeness we also include in the plots 0-th coefficients (often not informative).

We find that BERT contains considerable higher-frequency components, likely due to its non-autoregressive training; see also Wang & Chen (2020).

C Additional empirical results for Section 3

See Figure 19—Figure 23. Note that there are progressive cluster compactness changes across layers in BLOOM and Llama 2. It is likely that pretraining on heterogeneous datasets creates multiscale cluster structure. An investigation of this phenomenon is left as future work.

Measuring cluster compactness. We define Σ_W and Σ_B as the within-cluster and between-cluster covariance matrix respectively, and we use $\mathrm{Tr}(\Sigma_B\Sigma_W^{-1})$ to measure how well the samples are separated into clusters (bigger value is better). To compare the performance of **cvec** and raw embeddings in the downstream clustering tasks, we calculate them based on four documents of OpenWebText. In Figure 24, Figure 25, and Figure 26, we show the first two principal components for the **cvec** (left) and raw embeddings (right), with samples from documents shown in different colors. Based on the metric of $\mathrm{Tr}(\Sigma_B\Sigma_W^{-1})$ in the title and the PCA plot, we can see that **cvec**-s are better separated than the raw embeddings, indicating that the removal of positional basis is good for clustering tasks.

D Additional empirical results for Section 4

D.1 On QK matrix decomposition and induction heads

On global mean vector. We show the QK matrix decomposition with global mean (Figure 28), and without global mean (Figure 27). Note that adding a constant to all entries of the QK matrix will not change the attention matrix, because softmax computes the ratio. We conclude that the global mean vector $\boldsymbol{\mu}$ has little effects on interpretations.

See Figure 29—Figure 35 for more QK plots at various layers and heads on GPT-2 and BLOOM model.

D.2 On attention weight matrix

See Figure 36—Figure 41 for more plots on rotated $\boldsymbol{W} = \boldsymbol{W}^q(\boldsymbol{W}^k)^\top/\sqrt{d_{\text{head}}}$ at various layers and heads for BERT and GPT-2 model.

E Additional empirical results for Section E

Full results on the Randomization experiment. We apply three noise levels (regular, partially random, fully random) in both of the training and inference process. We give the $3 \times 3 = 9$ results in Figure 42—Figure 50.

Addition experiment. We have manually generate the addition dataset for the carry and no-carry tasks, with training set and validation set containing 100K and 10K additions of length ranging from 5 to 10 respectively. The models achieve 100% and 72% accuracy on the validation set in the no-carry and carry experiments, respectively. However, the model does not generalize: they get 5.02% and 0.00% accuracy on 1K samples of additions with length from 1 to 4. See Figure 51—Figure 56 for the QK plots at various layers and heads under carry and no-carry addition tasks. Notice the

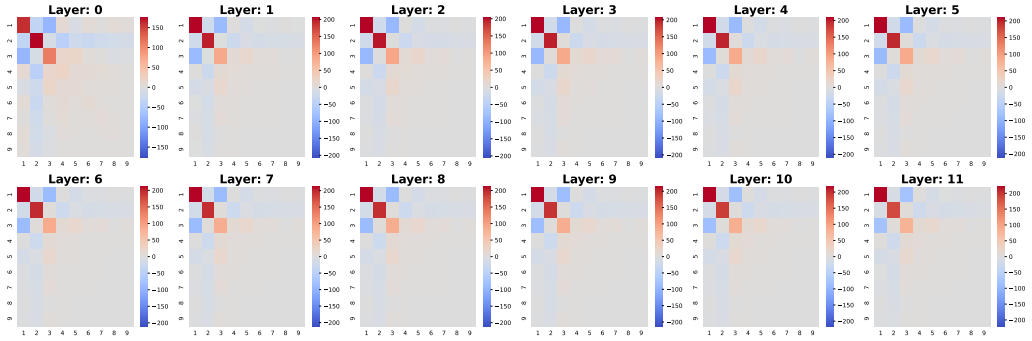

Figure 13: Fourier transformed positional basis; Openwebtext, GPT2

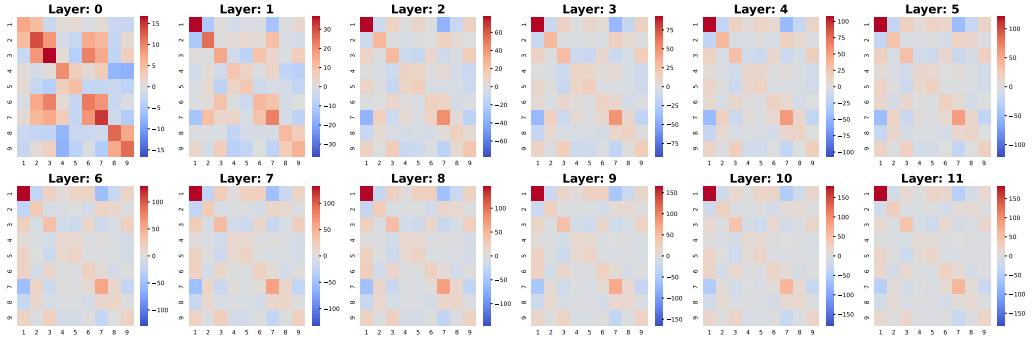

Figure 14: Fourier transformed positional basis; Openwebtext, BERT

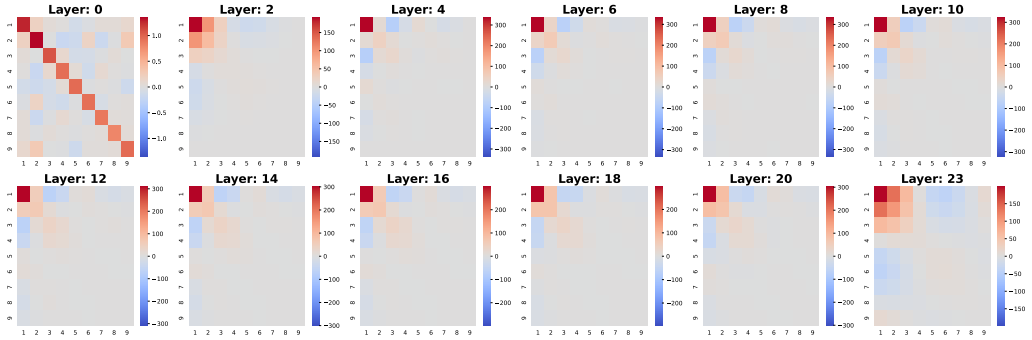

Figure 15: Fourier transformed positional basis; Openwebtext, BLOOM

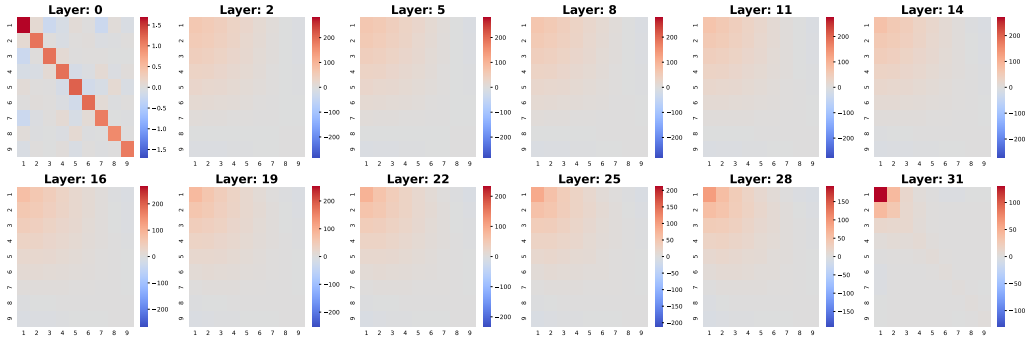

Figure 16: Fourier transformed positional basis; Openwebtext, Llama2

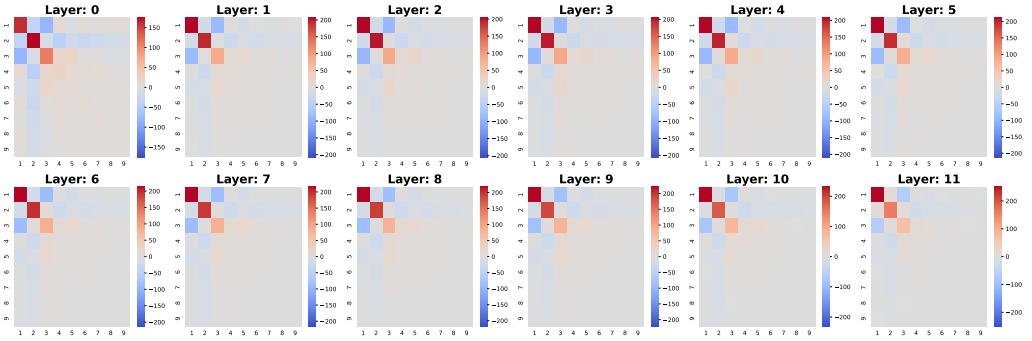

Figure 17: Fourier transformed positional basis; GitHub, GPT2

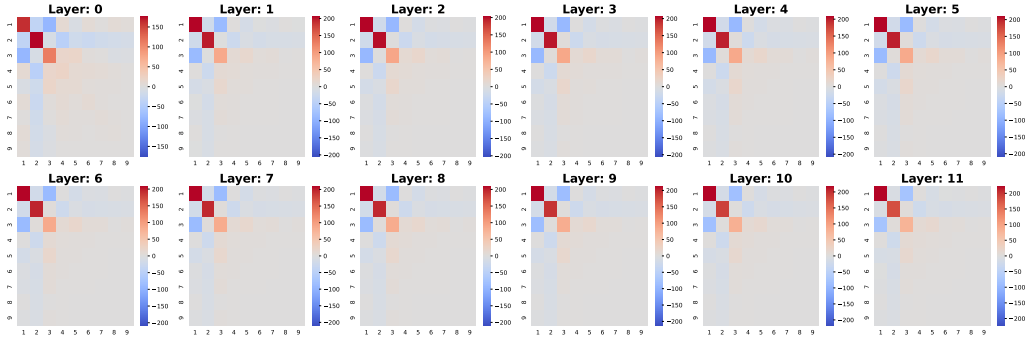

Figure 18: Fourier transformed positional basis; WikiText, GPT2

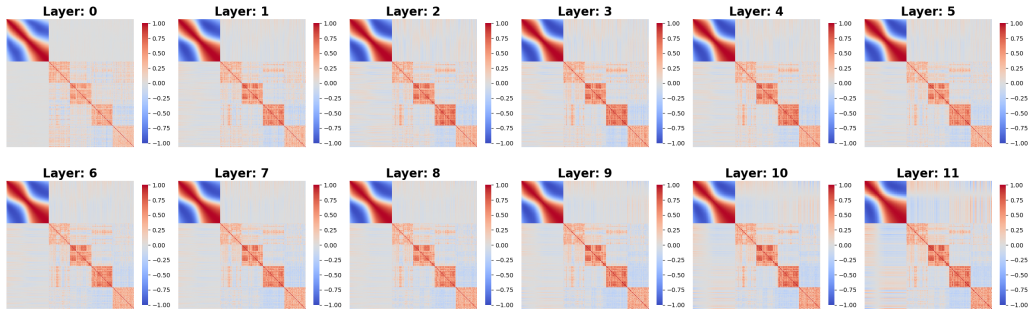

Figure 19: Gram matrix of positional basis and context basis; Openwebtext, GPT2

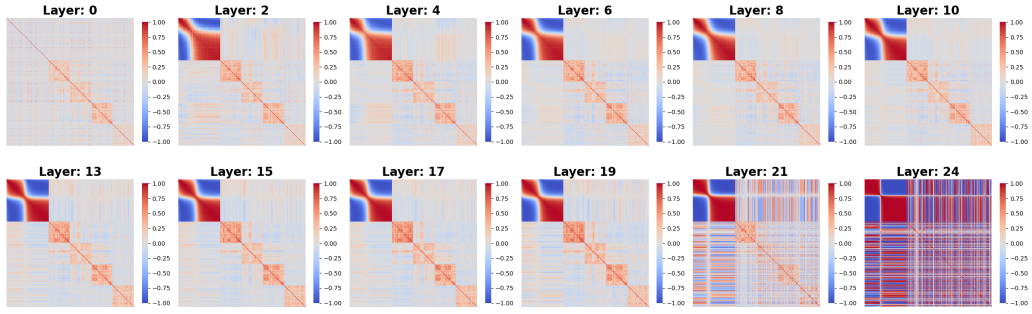

Figure 20: Gram matrix of positional basis and context basis; Openwebtext, BLOOM

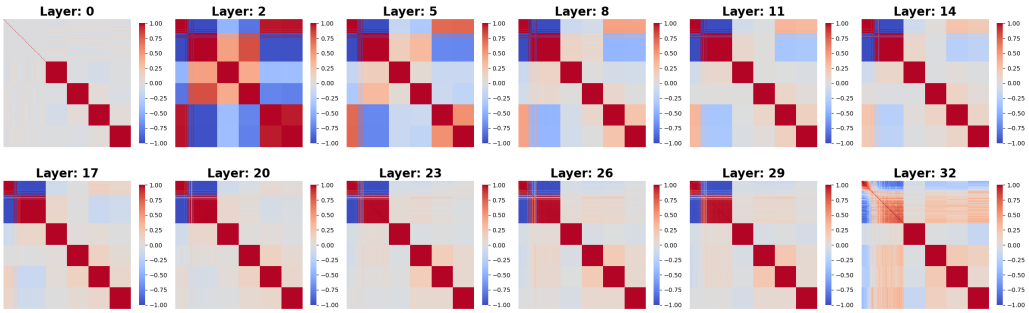

Figure 21: Gram matrix of positional basis and context basis; Openwebtext, Llama2

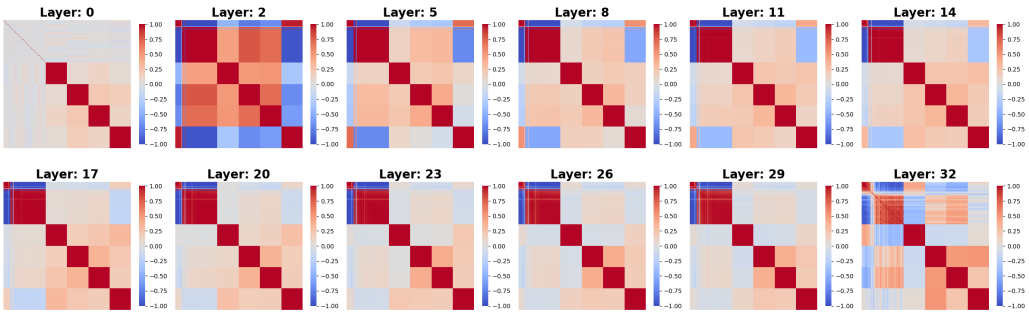

Figure 22: Gram matrix of positional basis and context basis; GitHub, Llama

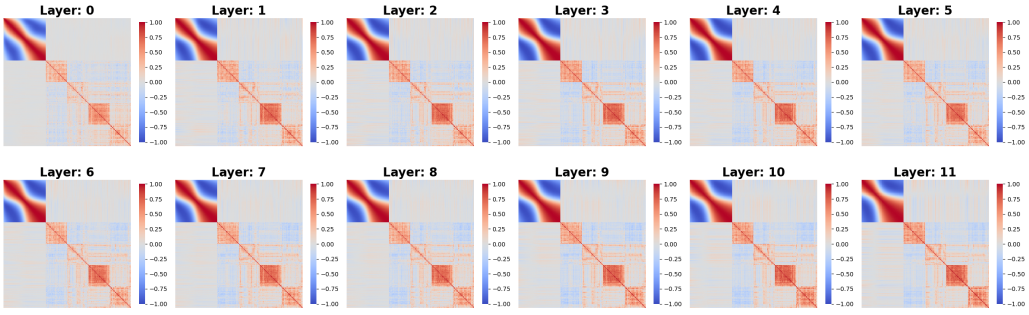

Figure 23: Gram matrix of positional basis and context basis; Wikitext, GPT2

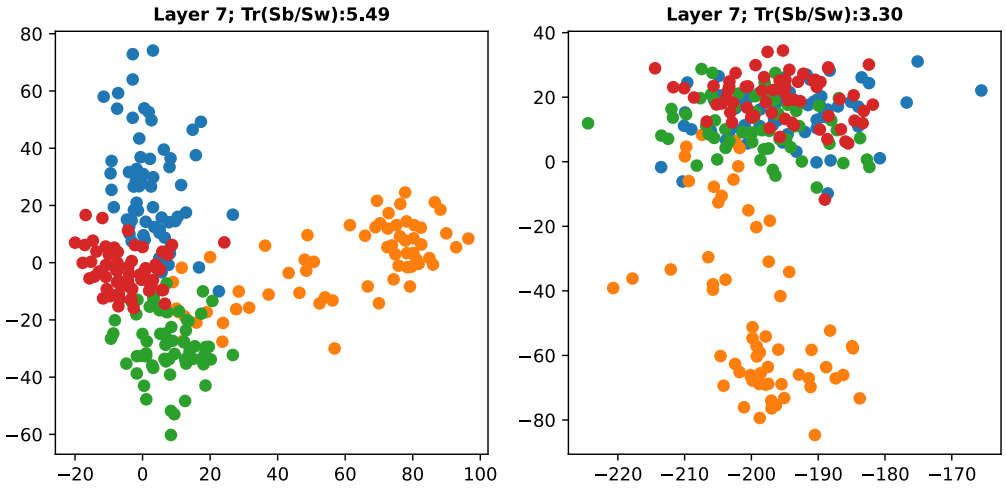

Figure 24: Top-2 principal components; GPT2, OpenWebText, L7. Left shows cvecs and right shows raw embeddings.

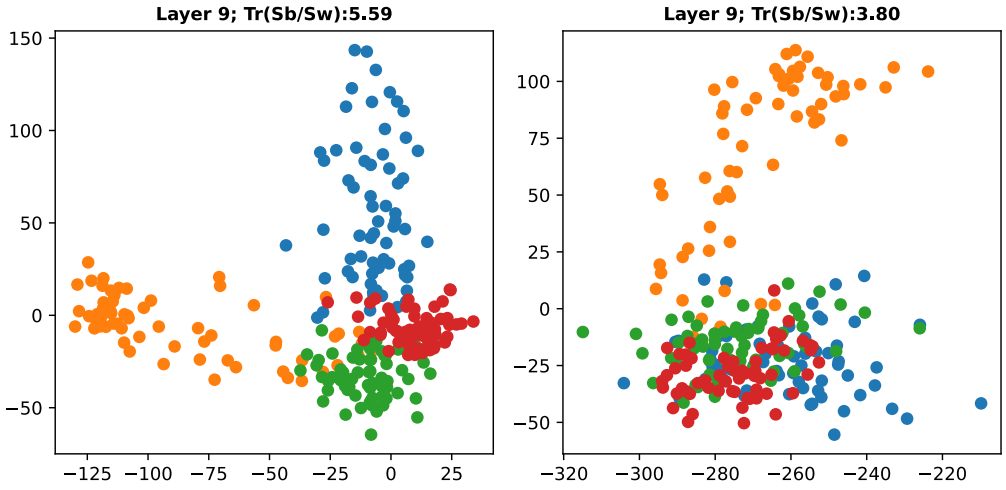

Figure 25: Top-2 principal components; GPT2, OpenWebText, L9. Left shows cvecs and right shows raw embeddings.

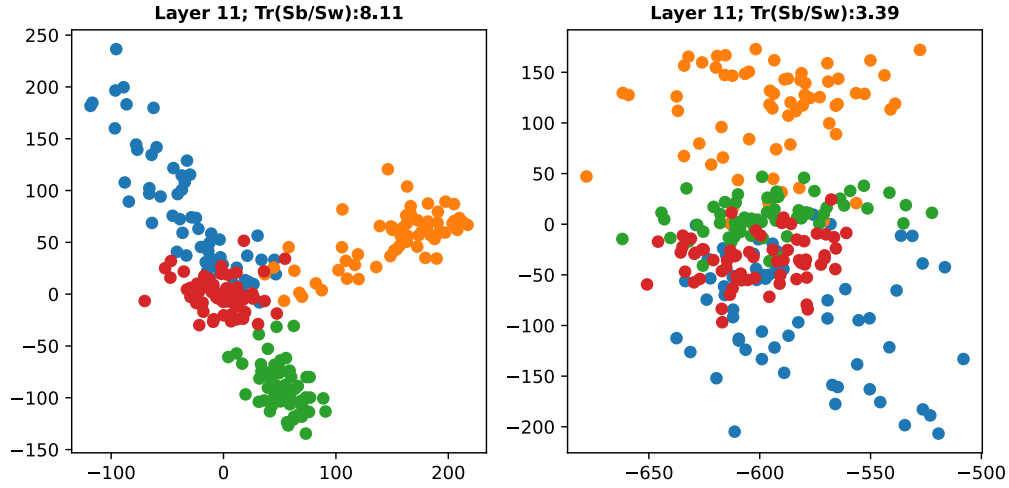

Figure 26: Top-2 principal components; GPT2, OpenWebText, L11. Left shows cvecs and right shows raw embeddings.

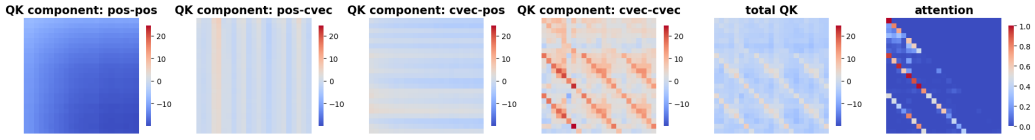

Figure 27: GPT2 L10H7 QK Decomposition; Global mean removed

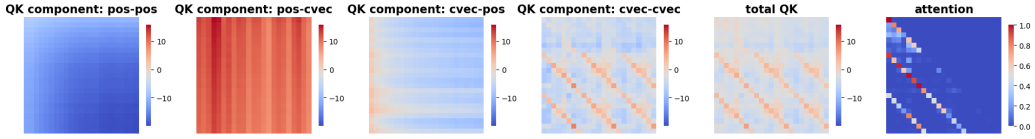

Figure 28: GPT2 L10H7 QK Decomposition; Global mean not removed

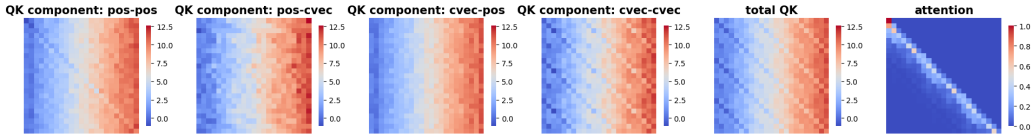

Figure 29: QK Decomposition; BLOOM L10H1

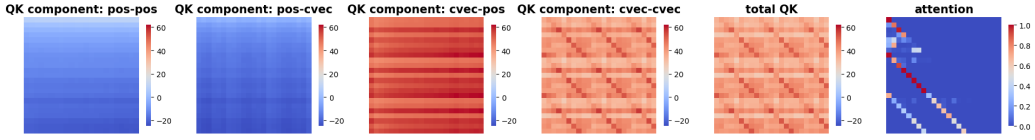

Figure 30: QK Decomposition; BLOOM L6H12

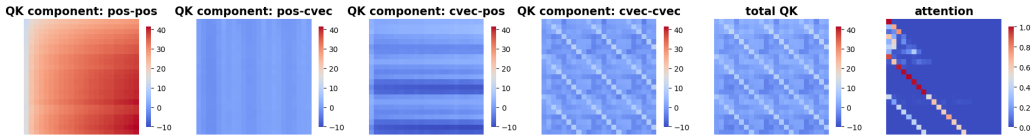

Figure 31: QK Decomposition; BLOOM L7H10

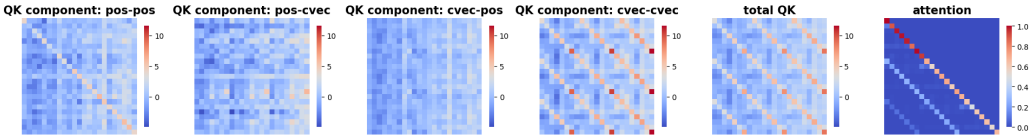

Figure 32: QK Decomposition; BLOOM L0H5

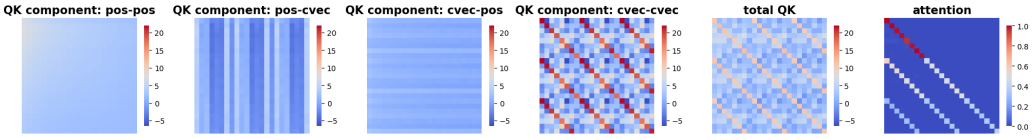

Figure 33: QK Decomposition; GPT2 L0H1

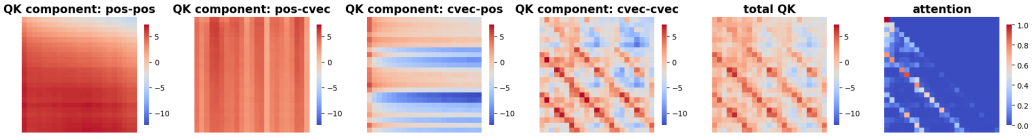

Figure 34: QK Decomposition; GPT2 L10H1

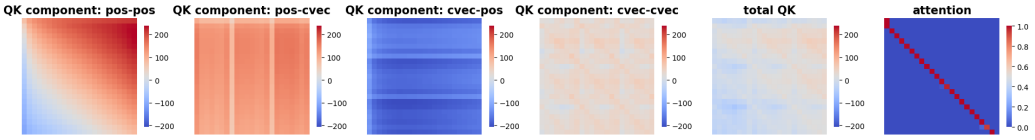

Figure 35: QK Decomposition; GPT2 L4H11

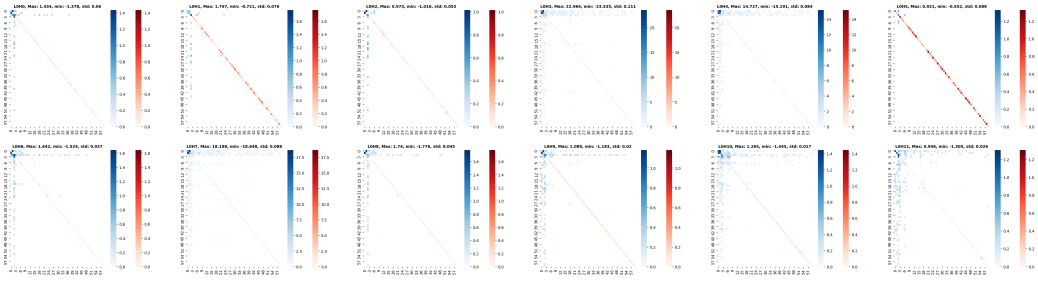

Figure 36: Dissecting attention weights, GPT2 L0

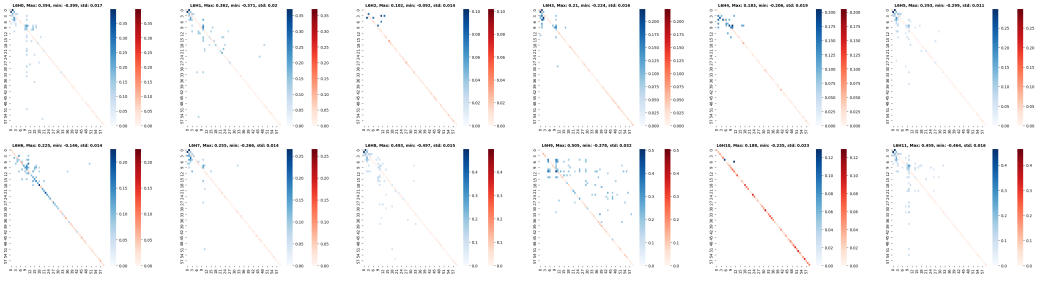

Figure 37: Dissecting attention weights, GPT2 L6

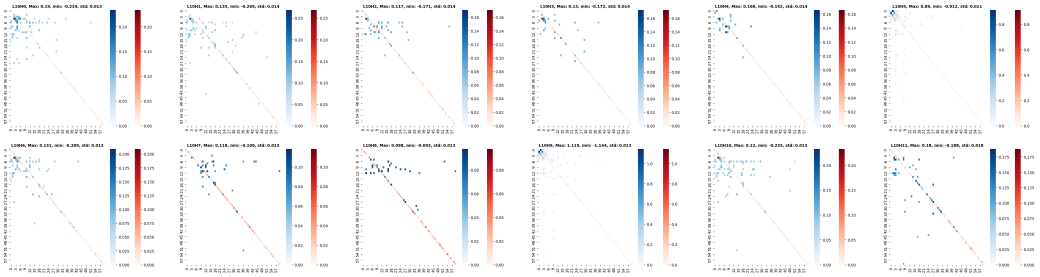

Figure 38: Dissecting attention weights, GPT2 L10

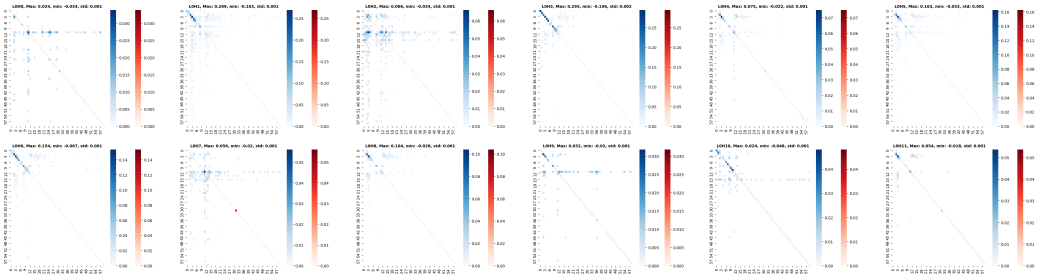

Figure 39: Dissecting attention weights, BERT L0

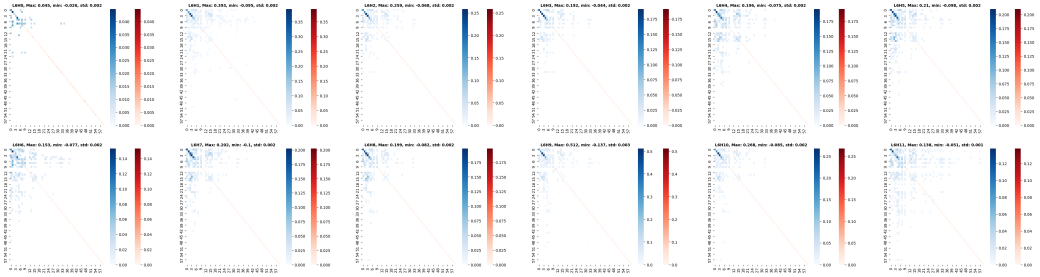

Figure 40: Dissecting attention weights, BERT L6

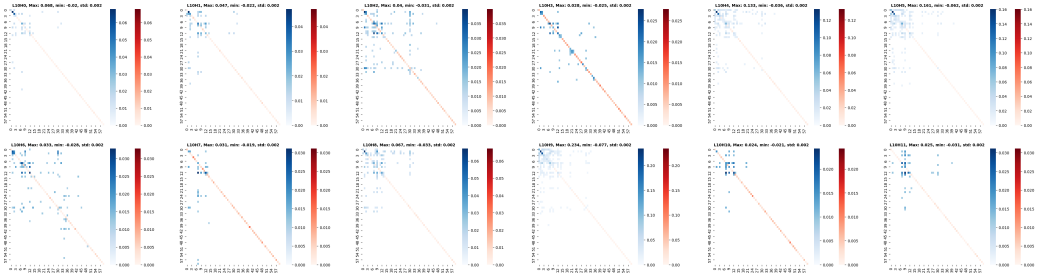

Figure 41: Dissecting attention weights, BERT L10

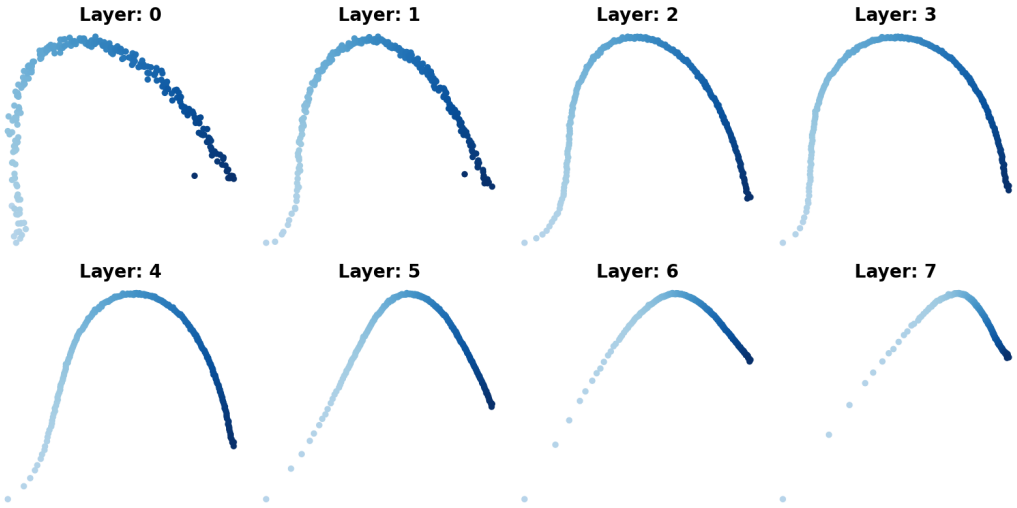

Figure 42: Training: regular; Inference: regular

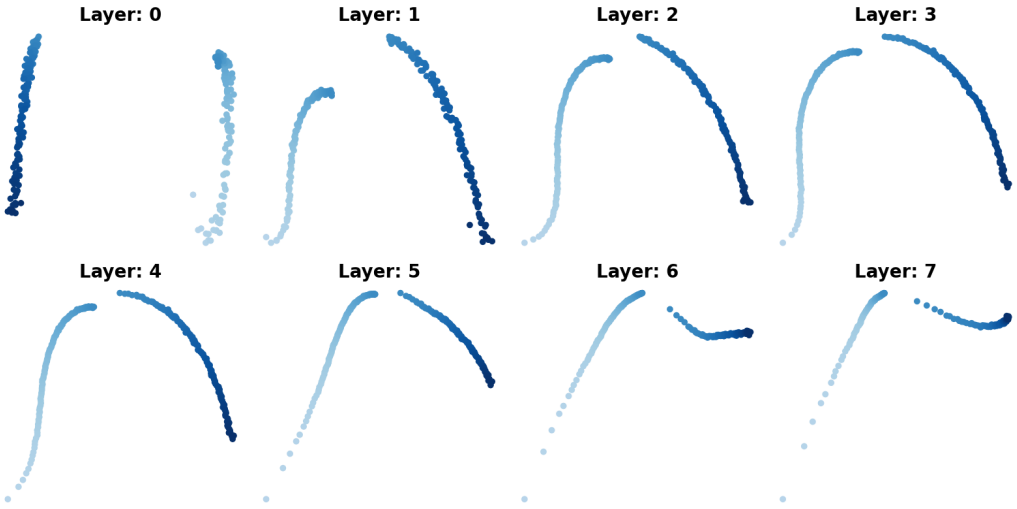

Figure 43: Training: regular; Inference: partially random

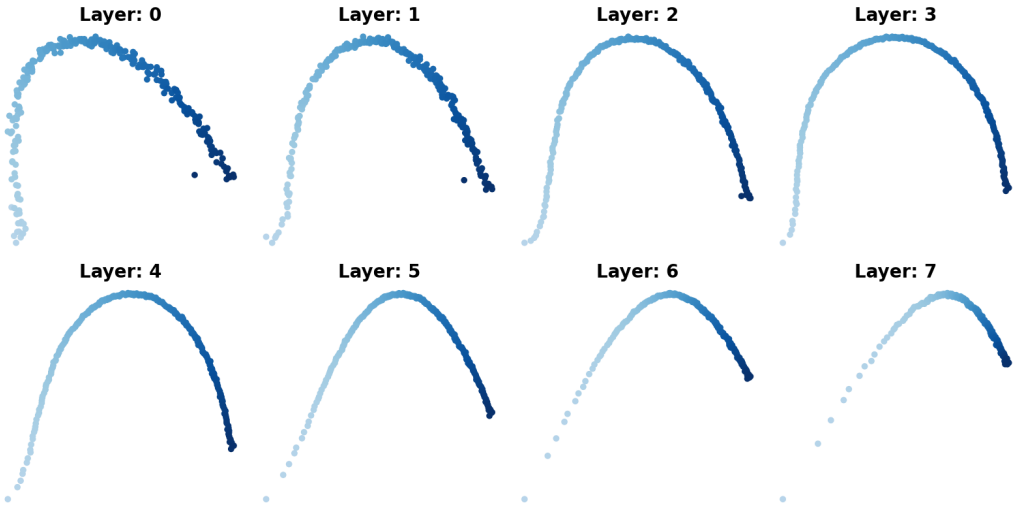

Figure 44: Training: regular; Inference: fully random

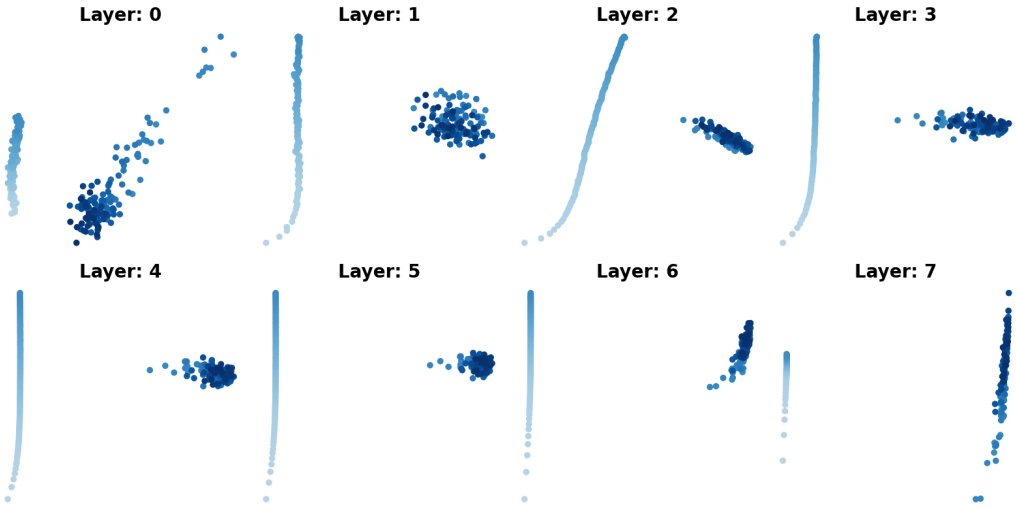

Figure 45: Training: partially random; Inference: regular

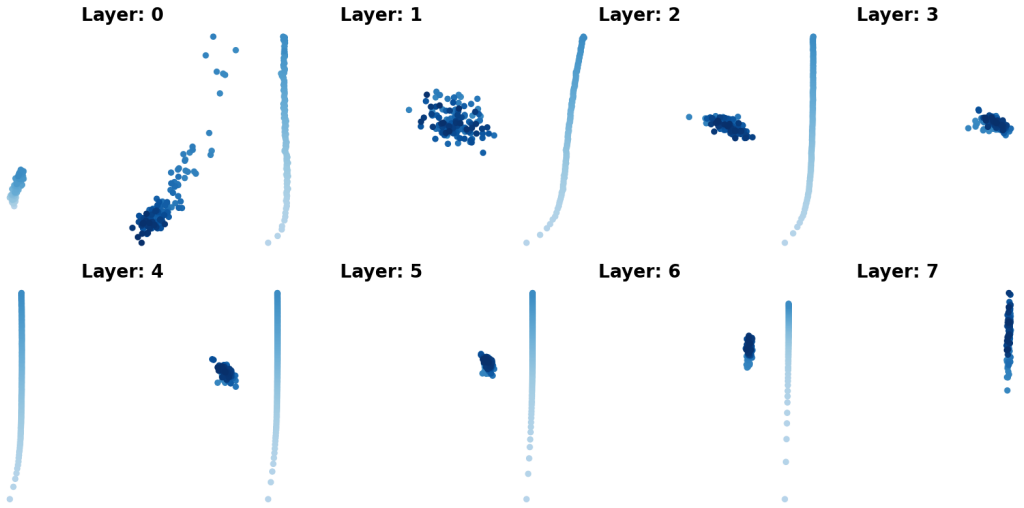

Figure 46: Training: partially random; Inference: partially random

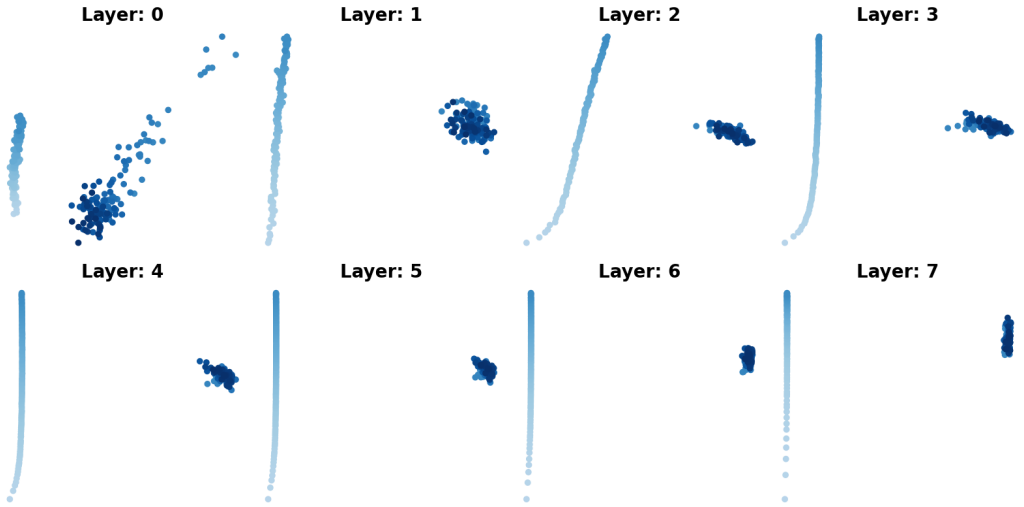

Figure 47: Training: partially random; Inference: fully random

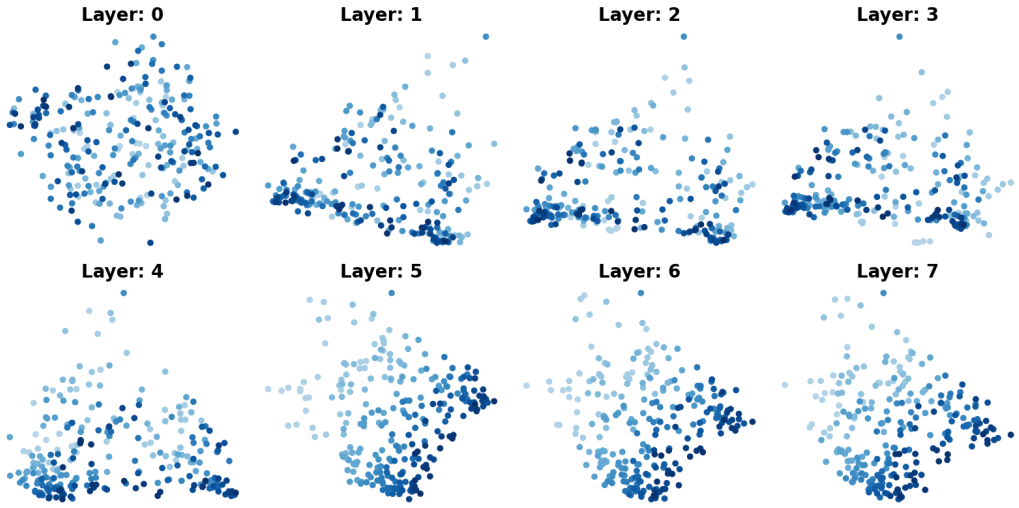

Figure 48: Training: fully random; Inference: regular

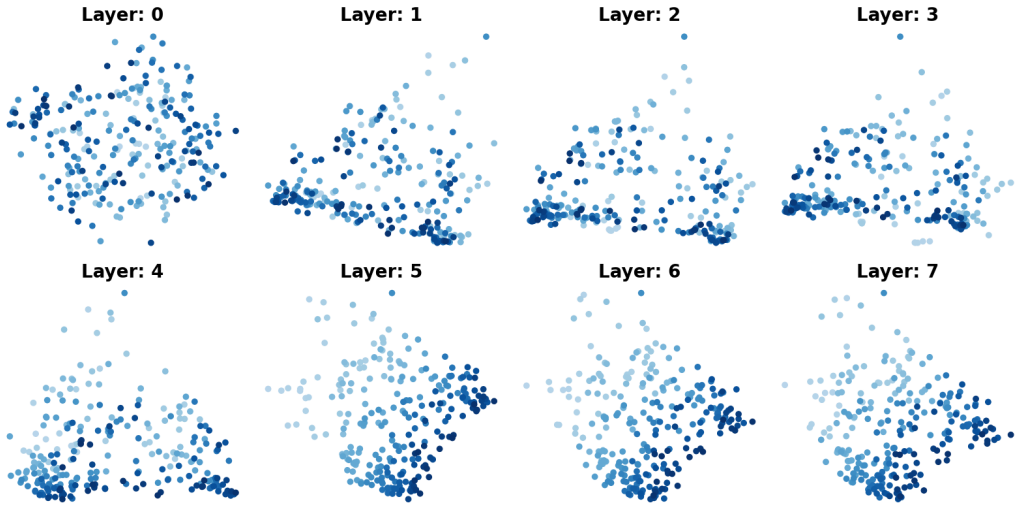

Figure 49: Training: fully random; Inference: partially random

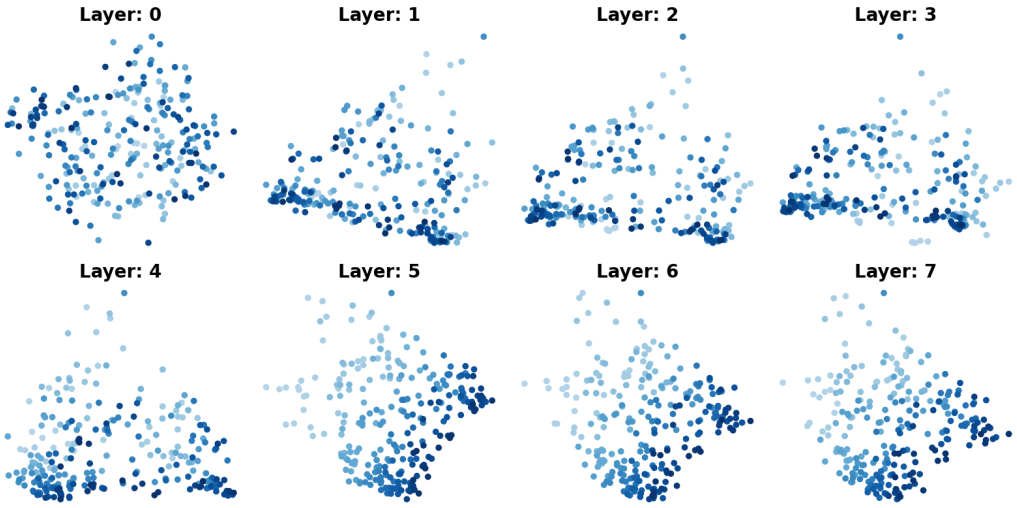

Figure 50: Training: fully random; Inference: fully random

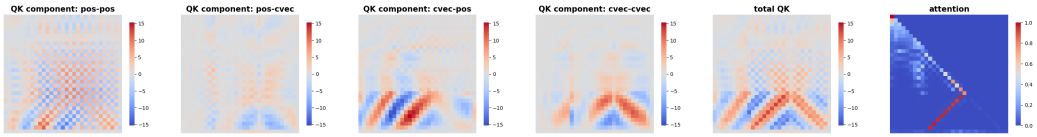

Figure 51: Addition without carry, L2H3

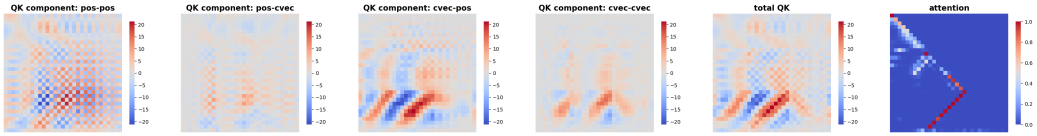

Figure 52: Addition without carry, L3H0

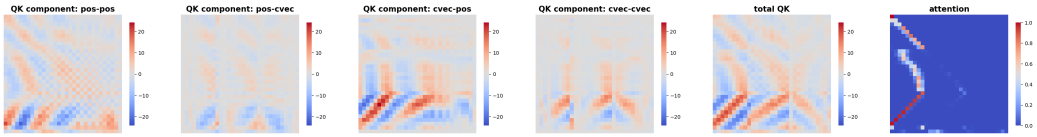

Figure 53: Addition without carry, L3H1

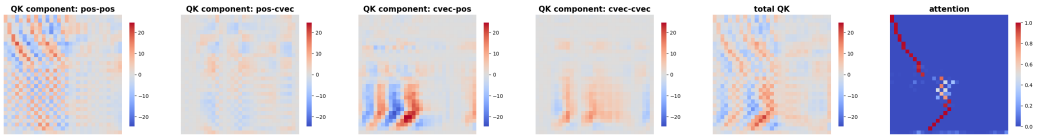

Figure 54: Addition with carry, L1H0

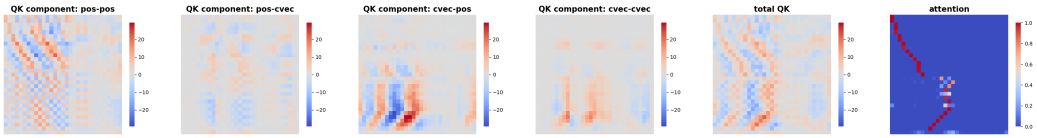

Figure 55: Addition with carry, L1H4

unsmoothness in the positional basis gram matrix that are pervasive across different layers, heads, and tasks.

F PROOFS FOR THEORETICAL RESULTS

We introduce some additional notations. Denote the indicator function by $\mathbf{1}$. We denote by $\mathbf{1}_N$ the vector $(1, 1, \ldots, 1)^\top \in \mathbb{R}^N$. For a complex matrix $\boldsymbol{A}$, we denote the conjugate transpose by $\boldsymbol{A}^*$. For convenience, for a matrix $\boldsymbol{A}$, we will write $\Delta\boldsymbol{A}$ instead of $\Delta^{(1,1)}\boldsymbol{A}$. For a vector $\boldsymbol{x} \in \mathbb{C}^N$, we also write $\Delta\boldsymbol{x}$ to denote the finite difference vector $N \cdot (x_1 - x_0, x_2 - x_1, \ldots, x_N - x_{N-1})^\top$ (where $x_N = x_0$). We will say that a Hermitian matrix $\boldsymbol{A} \in \mathbb{C}^{N\times N}$ is positive semidefinite (PSD) if and only if $\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x} \geq 0$ for every $\boldsymbol{x} \in \mathbb{C}^N$. Denote by $\Re(x)$ the real part of a complex number $\boldsymbol{x} \in \mathbb{C}$.

F.1 PROOF OF THEOREM 1

In this subsection, we denote a generic dimension by N and $\omega = \exp(-2\pi i/N)$. We need some standard definitions and properties; see Broughton & Bryan (2018) for example.

The discrete Fourier transform (DFT) matrix $\boldsymbol{F} \in \mathbb{C}^{N\times N}$ is given by $F_{tt'} = \omega^{(t-1)(t'-1)}$ for $1 \leq t, t' \leq N$. The inverse discrete Fourier transform (IDFT) matrix is $N^{-1}\boldsymbol{F}^*$. Both the DFT matrix and the IDFT matrix are symmetric (not Hermitian). Sometimes we prefer to write $\boldsymbol{F}^\top$ instead of $\boldsymbol{F}$ simply for formality. For a generic matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$, we denote by $\hat{\boldsymbol{A}} \in \mathbb{C}^{N\times N}$ the matrix after its 2-d DFT. It satisfies

$$\begin{aligned} \hat{\boldsymbol{A}} &= \boldsymbol{F}\boldsymbol{A}\boldsymbol{F}^\top, \\ \boldsymbol{A} &= N^{-2}\boldsymbol{F}^*\hat{\boldsymbol{A}}(\boldsymbol{F}^*)^\top. \end{aligned} \tag{13}$$

The following simple lemma is a consequence of integration-by-parts for the discrete version. For completeness we include a proof.

Lemma 1. *Let $\boldsymbol{x} \in \mathbb{R}^N$ be a vector, and $\hat{\boldsymbol{x}} = \boldsymbol{F}\boldsymbol{x}$ be its DFT. Then for $t = 1, \ldots, N$,*

$$\hat{x}_t = \gamma_t(\boldsymbol{F}\Delta\boldsymbol{x})_t + \mathbf{1}\{t = 1\} \cdot \sum_{t'=1}^{N} x_{t'}, \qquad \textit{where} \tag{14}$$

$$\gamma_t := N^{-1}\left(1 - \exp\left(\frac{-2\pi i(t-1)}{N}\right)\right)^{-1} \quad \textit{for } t > 1 \textit{ and } \gamma_1 := 1. \tag{15}$$

Proof. If $t = 1$, then $(\boldsymbol{F}\boldsymbol{x})_t = \sum_{t'=1}^N x_{t'}$, and $(\boldsymbol{F}\Delta\boldsymbol{x})_t = N\sum_{t'=1}^N (x_{t'} - x_{t'-1}) = 0$. For $t \neq 1$,

$$\begin{aligned} (\boldsymbol{F}\Delta\boldsymbol{x})_t &= N\sum_{t'=1}^{N} \omega^{(t-1)(t'-1)}(x_{t'} - x_{t'-1}) = N\sum_{t'=1}^{N} \big(\omega^{(t-1)(t'-1)} - \omega^{(t-1)t'}\big)x_{t'} \\ &= N\big(1 - \omega^{(t-1)}\big)\sum_{t'=1}^{N} \omega^{(t-1)(t'-1)}x_{t'} = \gamma_t^{-1}(\boldsymbol{F}\boldsymbol{x})_t. \end{aligned}$$

This shows $\hat{x}_t = \gamma_t(\boldsymbol{F}\Delta\boldsymbol{x})_t$ for $t \neq 1$ and completes the proof. □

A simple bound on the modulus $|\gamma_t|$ is given by the following lemma.

Lemma 2. *Let γ_t be defined by Equation 15. For positive integer $1 < t \leq N/2$, we have*

$$|\gamma_{N-t+2}| = |\gamma_t| \leq \frac{1}{8(t-1)}\,.$$

Proof. The equality part is obvious. For any $\theta \in (-\pi, \pi)$, we have

$$|1 - \exp(i\theta)|^2 = (1 - \cos\theta)^2 + \sin^2\theta = 4\sin^2(\theta/2) = 4\sin^2(|\theta|/2)$$

Since $\sin\theta/\theta$ is monotone decreasing in $(0, \pi/2)$, we have $\sin(|\theta|/2) \geq \sin(\pi/2)|\theta|/(\pi/2)$, Thus,

$$|1 - \exp(i\theta)| \geq 4|\theta|/\pi\,.$$

Setting $\theta = -2\pi(t-1)/N$, we obtain the desired upper bound on $|\gamma_t|$. □

Denote $\boldsymbol{\Gamma} = \mathrm{diag}\{\gamma_1, \ldots, \gamma_N\}$. By Lemma 1, for any vector $\boldsymbol{x} \in \mathbb{R}^N$,

$$\boldsymbol{F}\boldsymbol{x} = \boldsymbol{\Gamma}\boldsymbol{F}\Delta\boldsymbol{x} + \begin{pmatrix} \boldsymbol{x}^\top \mathbf{1}_N \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Below we will assume that the generic matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ is symmetric and satisfies $\boldsymbol{A}\mathbf{1}_N = \mathbf{0}$. Observe that

$$\begin{aligned} \boldsymbol{F}\boldsymbol{A} &= [\boldsymbol{F}\boldsymbol{A}_{:,1}, \ldots, \boldsymbol{F}\boldsymbol{A}_{:,N}] = \boldsymbol{\Gamma}\boldsymbol{F}[\Delta\boldsymbol{A}_{:,1}, \ldots, \Delta\boldsymbol{A}_{:,N}], \\ \boldsymbol{F}\boldsymbol{A}\boldsymbol{F}^\top &= \boldsymbol{\Gamma}\boldsymbol{F}(\Delta\boldsymbol{A})\boldsymbol{F}^\top\boldsymbol{\Gamma} \end{aligned}$$

where we used $\boldsymbol{A}\mathbf{1}_N = \mathbf{0}$ and $(\Delta\boldsymbol{A})\mathbf{1}_N = \mathbf{0}$. Repeating the second equality m times, we obtain

$$\hat{\boldsymbol{A}} = \boldsymbol{F}\boldsymbol{A}\boldsymbol{F}^\top = \boldsymbol{\Gamma}^m\boldsymbol{F}(\Delta^{(m,m)}\boldsymbol{A})\boldsymbol{F}^\top\boldsymbol{\Gamma}^m.$$

Now fix a generic nonempty index sets $\mathcal{I} \subset \{1, 2, \ldots, N\}$ and denote $\mathcal{J} = \{1, \ldots, N\}\backslash\mathcal{I}$. Consider the block matrix form of $\hat{\boldsymbol{A}}$:

$$\hat{\boldsymbol{A}} = \begin{pmatrix} \hat{\boldsymbol{A}}_{\mathcal{I},\mathcal{I}} & \hat{\boldsymbol{A}}_{\mathcal{I},\mathcal{J}} \\ \hat{\boldsymbol{A}}_{\mathcal{J},\mathcal{I}} & \hat{\boldsymbol{A}}_{\mathcal{J},\mathcal{J}} \end{pmatrix}.$$

Using the block matrix notation, we derive

$$\|\hat{\boldsymbol{A}}_{\mathcal{I},\mathcal{J}}\|_{\mathrm{op}} = \|\boldsymbol{\Gamma}^m_{\mathcal{I},\mathcal{I}}\big(\boldsymbol{F}(\Delta^{(m,m)}\boldsymbol{A})\boldsymbol{F}^\top\big)_{\mathcal{I},\mathcal{J}}\boldsymbol{\Gamma}^m_{\mathcal{J},\mathcal{J}}\|_{\mathrm{op}} \le \max_{t\in\mathcal{I},t'\in\mathcal{J}} |\gamma_t\gamma_{t'}|^m \|\boldsymbol{F}(\Delta^{(m,m)}\boldsymbol{A})\boldsymbol{F}^\top\|_{\mathrm{op}}.$$

Similar inequalities hold for other three blocks. By Lemma 2 we have $|\gamma_t| \le 1$ for all t. Adding the three inequalities that involve at least one index set $\mathcal{J}$, we get

$$\left\|\hat{\boldsymbol{A}} - \begin{pmatrix} \hat{\boldsymbol{A}}_{\mathcal{I},\mathcal{I}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\right\|_{\mathrm{op}} \le 3\max_{t\in\mathcal{J}} |\gamma_t|^m \|\boldsymbol{F}(\Delta^{(m,m)}\boldsymbol{A})\boldsymbol{F}^\top\|_{\mathrm{op}}.$$

Since $\boldsymbol{F}\boldsymbol{F}^* = N$, we have $\|\boldsymbol{F}\|_{\mathrm{op}} = \|\boldsymbol{F}\boldsymbol{F}^*\|^{1/2}_{\mathrm{op}} = N^{1/2}$. Denoting

$$\boldsymbol{A}^{(\mathrm{res})} := N^{-2}\boldsymbol{F}^*\left[\hat{\boldsymbol{A}} - \begin{pmatrix} \hat{\boldsymbol{A}}_{\mathcal{I},\mathcal{I}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\right](\boldsymbol{F}^*)^\top,$$

we find

$$\|\boldsymbol{A}^{(\mathrm{res})}\|_{\mathrm{op}} \le 3\max_{t\in\mathcal{J}} |\gamma_t|^m N^{-2}\|\boldsymbol{F}\|^4_{\mathrm{op}}\|\Delta^{(m,m)}\boldsymbol{A}\|_{\mathrm{op}} \le 3\max_{t\in\mathcal{J}} |\gamma_t|^m N\|\Delta^{(m,m)}\boldsymbol{A}\|_{\max} \tag{16}$$

where the first inequality is due to Lemma 2 and the second inequality is due to the inequality between matrix operator norm and max norm.

To finish the proof, let us make some specification: we identify N with $\tilde{T}$ (namely $2T$), identify $\boldsymbol{A}$ with $\tilde{\boldsymbol{G}}$, and identify $\mathcal{I} = \{1, \ldots, k\} \cup \{\tilde{T} - k + 1, \ldots, \tilde{T}\}$. These choices satisfy the requirement for $\boldsymbol{A}$ because $\tilde{\boldsymbol{G}}$ is symmetric and it satisfies $\tilde{\boldsymbol{G}}\mathbf{1}_{\tilde{T}} = 2\boldsymbol{G}\mathbf{1}_T = \mathbf{0}$ due to the assumption $\mathbf{pos}_1 + \ldots + \mathbf{pos}_T = \mathbf{0}$.

By Lemma 2, $\max_{t\in\mathcal{J}} |\gamma_t| \le 1/(8k)$, so Equation 16 gives

$$\|\boldsymbol{A}^{(\mathrm{res})}\|_{\mathrm{op}} \le 6(8k)^{-m}T\|\Delta^{(m,m)}\tilde{\boldsymbol{G}}\|_{\max}$$

By the definition of $\boldsymbol{A}^{(\mathrm{res})}$ and the identities in 13,

$$\tilde{\boldsymbol{G}} = \boldsymbol{A}^{(\mathrm{res})} + N^{-2}\boldsymbol{F}^*\begin{pmatrix} \hat{\boldsymbol{A}}_{\mathcal{I},\mathcal{I}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}(\boldsymbol{F}^*)^\top.$$

We make the following claim.

Lemma 3. *There exists $\boldsymbol{B} \in \mathbb{R}^{k\times k}$ such that*

$$N^{-2}\left[\boldsymbol{F}^*\begin{pmatrix} \hat{\boldsymbol{A}}_{\mathcal{I},\mathcal{I}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}(\boldsymbol{F}^*)^\top\right]_{1:T,1:T} = \boldsymbol{F}_{\le k}\boldsymbol{B}(\boldsymbol{F}_{\le k}\boldsymbol{B})^\top \tag{17}$$

While the DFT matrix is complex, the above lemma claims that the left-hand side is a Gram matrix of real low-frequency vectors. Once this lemma is proved, we can combine this lemma with Equation 16 and $\boldsymbol{G} = \tilde{\boldsymbol{G}}_{1:T,1:T}$ to obtain the desired inequality 6 in Theorem 1.

Proof of Lemma 3. Recall $\omega = \exp(-2\pi i/2T)$. We further introduce some notations. Denote $\mathcal{I}_1 = \{1, \dots, k\}$ and $\mathcal{I}_2 = \{T-k+1, \dots, T\}$ so that $\mathcal{I} = \mathcal{I}_1 \cup \mathcal{I}_2$. Let $q_t = \omega^{t-1}$ for positive integer t, matrix $\boldsymbol{Q} \in \mathbb{R}^{T\times k}$ and matrix $\boldsymbol{D} \in \mathbb{R}^{k\times k}$ be given by

$$Q_{t,s} = \Re(q_t^{s-0.5}), \qquad \text{where } t \le T, s \le k, \qquad \boldsymbol{D} = \operatorname{diag}(q_1^{1/2}, \dots, q_k^{1/2})$$

For $t, t' \in \mathcal{I}$,

$$\begin{aligned}
\hat{A}_{t,t'} &= (\boldsymbol{F}\boldsymbol{A}\boldsymbol{F}^T)_{t,t'} \\
&= \sum_{s,s'=1}^{T} F_{t,s}G_{s,s'}F_{t',s'} + \sum_{s,s'=1}^{T} F_{t,2T+1-s}G_{s,s'}F_{t',s'} \\
&+ \sum_{s,s'=1}^{T} F_{t,s}G_{s,s'}F_{t',2T+1-s'} + \sum_{s,s'=1}^{T} F_{t,2T+1-s}G_{s,s'}F_{t',2T+1-s'} \\
&= q_t^{1/2}q_{t'}^{1/2} \sum_{s,s'=1}^{T} G_{s,s'} \left(q_t^{s-0.5}q_{t'}^{s'-0.5} + q_t^{-s+0.5}q_{t'}^{s'-0.5} + q_t^{s-0.5}q_{t'}^{-s'+0.5} + q_t^{-s+0.5}q_{t'}^{-s'+0.5} \right) \\
&= 4q_t^{1/2}q_{t'}^{1/2} \sum_{s,s'=1}^{T} G_{s,s'} \Re(q_t^{s-1/2}) \Re(q_{t'}^{s'-1/2}) \,.
\end{aligned}$$

If $t, t' \in \mathcal{I}_1$, the above equality leads to

$$\hat{\boldsymbol{A}}_{\mathcal{I}_1,\mathcal{I}_1} = 4\boldsymbol{D}\boldsymbol{Q}^\top \boldsymbol{G}\boldsymbol{Q}\boldsymbol{D};$$

and more generally $\hat{\boldsymbol{A}}_{\mathcal{I},\mathcal{I}}$ is given by symmetrically extending $\hat{\boldsymbol{A}}_{\mathcal{I}_1,\mathcal{I}_1}$ as in the definition of $\tilde{\boldsymbol{G}}$. Since $4\boldsymbol{Q}^\top\boldsymbol{G}\boldsymbol{Q}$ is a PSD, we can find $\boldsymbol{B}_0 \in \mathbb{R}^{k\times k}$ such that

$$4\boldsymbol{Q}^\top\boldsymbol{G}\boldsymbol{Q} = \boldsymbol{B}_0\boldsymbol{B}_0^\top \,.$$

We want to simplify $\boldsymbol{F}^*_{:,\mathcal{I}}\hat{\boldsymbol{A}}_{\mathcal{I},\mathcal{I}}\boldsymbol{F}^*_{\mathcal{I},:}$, namely

$$\boldsymbol{F}^*_{:,\mathcal{I}_1}\hat{\boldsymbol{A}}_{\mathcal{I}_1,\mathcal{I}_1}\boldsymbol{F}^*_{\mathcal{I}_1,:} + \boldsymbol{F}^*_{:,\mathcal{I}_1}\hat{\boldsymbol{A}}_{\mathcal{I}_1,\mathcal{I}_2}\boldsymbol{F}^*_{\mathcal{I}_2,:} + \boldsymbol{F}^*_{:,\mathcal{I}_2}\hat{\boldsymbol{A}}_{\mathcal{I}_2,\mathcal{I}_1}\boldsymbol{F}^*_{\mathcal{I}_1,:} + \boldsymbol{F}^*_{:,\mathcal{I}_2}\hat{\boldsymbol{A}}_{\mathcal{I}_2,\mathcal{I}_2}\boldsymbol{F}^*_{\mathcal{I}_2,:} \tag{18}$$

For any $t, t' \in \mathcal{I}_1$,

$$(\boldsymbol{F}^*)_{t,\mathcal{I}_1}\hat{\boldsymbol{A}}_{\mathcal{I}_1,\mathcal{I}_1}(\boldsymbol{F}^*)_{\mathcal{I}_1,t'} = (\boldsymbol{F}^*)_{t,\mathcal{I}_1}\boldsymbol{D}\boldsymbol{B}_0\boldsymbol{B}_0^\top\boldsymbol{D}(\boldsymbol{F}^*)_{\mathcal{I}_1,t'}$$

and similar equations hold for other three cases. Observe that

$$(\boldsymbol{F}^*)_{t,\mathcal{I}_1}\boldsymbol{D} = (\bar{q}_t^{1-0.5}, \dots, \bar{q}_t^{k-0.5}), \qquad (\boldsymbol{F}^*)_{t,\mathcal{I}_2}\boldsymbol{D} = (\bar{q}_t^{2T-0.5}, \dots, \bar{q}_t^{2T-k+0.5}) \,.$$

When we add the four terms in 18, the imaginary part cancels out. Thus,

$$\boldsymbol{F}^*_{t,\mathcal{I}}\hat{\boldsymbol{A}}_{\mathcal{I},\mathcal{I}}\boldsymbol{F}^*_{\mathcal{I},t'} = 4\left(\Re(\bar{q}_t^{1-0.5}), \dots, \Re(\bar{q}_t^{k-0.5})\right)\boldsymbol{B}_0\boldsymbol{B}_0^\top\left(\Re(\bar{q}_t^{1-0.5}), \dots, \Re(\bar{q}_t^{k-0.5})\right)^\top \,.$$

Note that $\Re(\bar{q}_t^{s-0.5}) = \cos(\pi(t-1)(s-0.5)/T) = (\boldsymbol{F}_{\le k})_{t,s}$. Expressing $\boldsymbol{F}^*_{:,\mathcal{I}}\hat{\boldsymbol{A}}_{\mathcal{I},\mathcal{I}}\boldsymbol{F}^*_{\mathcal{I},:}$ in the matrix form and denote $\boldsymbol{B} = \boldsymbol{B}_0/N$, we find that Equation 17 holds. □

F.2 Proof of Theorem 2

First we note that

$$K_{\boldsymbol{W}}(\boldsymbol{x}^q, \boldsymbol{x}^k) = K_{\boldsymbol{W}_{11}}(\boldsymbol{x}^q, \boldsymbol{x}^k)K_{\boldsymbol{W}_{12}}(\boldsymbol{x}^q, \boldsymbol{x}^k)K_{\boldsymbol{W}_{21}}(\boldsymbol{x}^q, \boldsymbol{x}^k)K_{\boldsymbol{W}_{22}}(\boldsymbol{x}^q, \boldsymbol{x}^k) \,. \tag{19}$$

We will prove that

$$K_{\boldsymbol{W}_{11}}(\boldsymbol{x}^q, \boldsymbol{x}^k) = \big(1 + O(\text{incoh})\big) \cdot K_{\boldsymbol{W}_{11}}(\boldsymbol{c}^q, \boldsymbol{c}^k) \,. \tag{20}$$

To prove this, it suffices to show that

$$\max\left\{K_{\boldsymbol{W}_{11}}(\boldsymbol{c}^q,\boldsymbol{t}^k), K_{\boldsymbol{W}_{11}}(\boldsymbol{t}^q,\boldsymbol{c}^k), K_{\boldsymbol{W}_{11}}(\boldsymbol{t}^q,\boldsymbol{t}^k)\right\} = 1 + O(\mathrm{incoh})\,. \tag{21}$$

We decompose $\boldsymbol{W}_{11}$ as in Equation 11 and find

$$\log\left(K_{\boldsymbol{W}_{11}}(\boldsymbol{c}^q,\boldsymbol{t}^k)\right) = (\boldsymbol{c}^q)^\top \boldsymbol{W}_{11}\boldsymbol{t}^k = \sum_{k=1}^{s} a_k(\boldsymbol{u}_k^\top \boldsymbol{c}^q)(\boldsymbol{v}_k^\top \boldsymbol{t}^k)\,.$$

By mutual incoherence, $|\boldsymbol{v}_k^\top \boldsymbol{t}^k| \le \mathrm{incoh}$ since $\boldsymbol{v}_k \in \mathcal{B}_1$ and $\boldsymbol{t}^k \in \mathcal{B}_2$; and trivially $|\boldsymbol{u}_k^\top \boldsymbol{c}^k| \le 1$, so

$$\left|\log\left(K_{\boldsymbol{W}_{11}}(\boldsymbol{c}^q,\boldsymbol{t}^k)\right)\right| \le \sum_{k=1}^{s} |\boldsymbol{u}_k^\top \boldsymbol{c}^k| \cdot |\boldsymbol{v}_k^\top \boldsymbol{t}^q| \le s \cdot \mathrm{incoh}\,.$$

Since by assumption $s = O(1)$ and $\exp(\mathrm{incoh}) = 1 + O(\mathrm{incoh})$, we derive

$$K_{\boldsymbol{W}_{11}}(\boldsymbol{c}^q,\boldsymbol{t}^k) = 1 + O(\mathrm{incoh})\,.$$

The other two terms in Equation 21 follow a similar argument and thus are all bounded by $1 + O(\mathrm{incoh})$. We can prove similarly that

$$\begin{aligned}
K_{\boldsymbol{W}_{12}}(\boldsymbol{x}^q,\boldsymbol{x}^k) &= \big(1 + O(\mathrm{incoh})\big) \cdot K_{\boldsymbol{W}_{12}}(\boldsymbol{c}^q,\boldsymbol{t}^k)\,,\\
K_{\boldsymbol{W}_{21}}(\boldsymbol{x}^q,\boldsymbol{x}^k) &= \big(1 + O(\mathrm{incoh})\big) \cdot K_{\boldsymbol{W}_{21}}(\boldsymbol{t}^q,\boldsymbol{c}^k)\,,\\
K_{\boldsymbol{W}_{21}}(\boldsymbol{x}^q,\boldsymbol{x}^k) &= \big(1 + O(\mathrm{incoh})\big) \cdot K_{\boldsymbol{W}_{22}}(\boldsymbol{t}^q,\boldsymbol{t}^k)\,.
\end{aligned}$$

and together with Equation 20 and Equation 19, this leads to the desired Equation 12.

Below we prove the "moreover" part. By standard properties of independent subgaussian random variables (Vershynin, 2018, Sect. 2), $(\boldsymbol{c}^q)^\top \mathbf{Z}_{11}\boldsymbol{t}^k$ is still a subgaussian random variable, and with probability at least $1 - O(\exp(-\mathrm{incoh}^2 \cdot d))$, for certain constant $C > 0$,

$$\left|\log\left(K_{\boldsymbol{W}_{11}+\mathbf{Z}/\sqrt{d}}(\boldsymbol{c}^q,\boldsymbol{t}^k)\right)\right| \le s \cdot \mathrm{incoh} + C\mathrm{incoh} = O(\mathrm{incoh}).$$

Similar high-probability bounds hold for other terms. By the union bound over all possible choice of vectors in $\mathcal{B}_1^0$ and $\mathcal{B}_2^0$, we arrive at our claim.

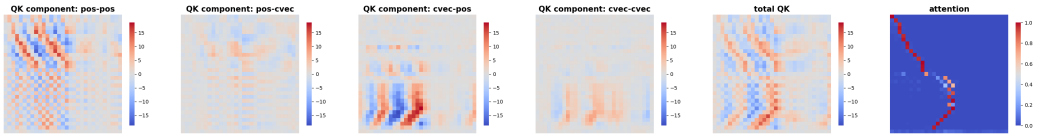

Figure 56: Addition with carry, L1H7