Norm of Mean Contextualized Embeddings Determines their Variance

Hiroaki Yamagiwa[1] **Hidetoshi Shimodaira**[1,2]
[1]Kyoto University [2]RIKEN
hiroaki.yamagiwa@sys.i.kyoto-u.ac.jp,shimo@i.kyoto-u.ac.jp

Abstract

Contextualized embeddings vary by context, even for the same token, and form a distribution in the embedding space. To analyze this distribution, we focus on the norm of the mean embedding and the variance of the embeddings. In this study, we first demonstrate that these values follow the well-known formula for variance in statistics and provide an efficient sequential computation method. Then, by observing embeddings from intermediate layers of several Transformer models, we found a strong trade-off relationship between the norm and the variance: as the mean embedding becomes closer to the origin, the variance increases. This trade-off is likely influenced by the layer normalization mechanism used in Transformer models. Furthermore, when the sets of token embeddings are treated as clusters, we show that the variance of the entire embedding set can theoretically be decomposed into the within-cluster variance and the between-cluster variance. We found experimentally that as the layers of Transformer models deepen, the embeddings move farther from the origin, the between-cluster variance relatively decreases, and the within-cluster variance relatively increases. These results are consistent with existing studies on the anisotropy of the embedding spaces across layers.

1 Introduction

Contextualized embedding is a method for dynamically computing the embeddings of tokens in a sentence. Unlike static embeddings such as Skip-gram (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), where a predefined embedding is assigned to each word, models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019b) compute contextualized embeddings based on the context, leading to superior performance in various downstream tasks. Even for the same token,

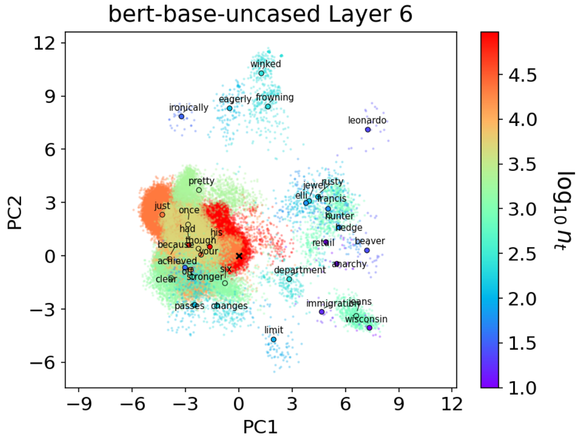


Figure 1: Scatter plots of PCA-transformed embeddings for the embedding sets X_t of selected tokens. The origin is indicated by $\times$. Tokens distributed near the origin exhibit larger variance, whereas tokens farther from the origin exhibit smaller variance. Embeddings are colored according to token frequency n_t.

the contextualized embeddings vary for sentences, creating a distribution in the embedding space.

Research has been done to explore the relationship between word frequency and contextualized embeddings. Wannasuphoprasit et al. (2023) showed a correlation between the frequency of a word and the mean norm of its BERT embeddings. Liang et al. (2021) found a negative correlation between the frequency and the norm of BERT embeddings. Zhou et al. (2021, 2022a) observed that higher frequency words tend to have a larger radius of the smallest enclosing sphere of their BERT embeddings. In particular, the larger radius value means the broader distribution of the embeddings. These studies reveal intriguing relationships between word frequency, the norm of embeddings, and the spread of their distribution.

Based on these existing studies, we analyze the distribution of embeddings using statistical measures computed from the first and second moments of the embedding components. Let $\boldsymbol{x}_{t,i}$ denote the

Our code is available at https://github.com/ymgw55/Norm-and-Variance.

Proceedings of the 31st International Conference on Computational Linguistics, pages 7778–7808
January 19–24, 2025.

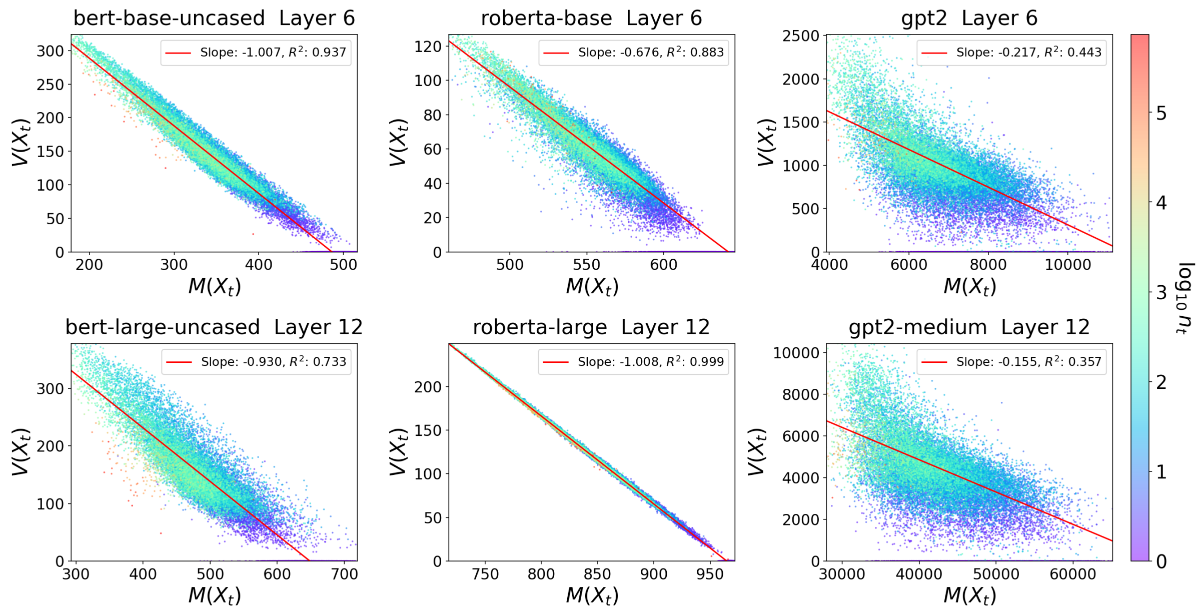


Figure 2: Scatter plots of $V(X_t)$ against $M(X_t)$ for the middle-layer embeddings of six models with regression lines, slopes, and coefficients of determination, R^2. A consistent trade-off between $M(X_t)$ and $V(X_t)$ is observed in the intermediate layer of each model. A summary for all the other layers can be found in Fig. 4. Only tokens with $1 \leq \log_{10} n_t \leq 5$ were used for regressions to reduce the influence of extreme values.

d-dimensional contextualized embedding for token type t in its i-th occurrence[1]. For the set of contextualized embeddings $X_t = \{\boldsymbol{x}_{t,1}, \boldsymbol{x}_{t,2}, \ldots\} \subset \mathbb{R}^d$ corresponding to token t, we focus on three values: **the mean squared norm** $Q(X_t)$, **the squared norm of the mean embedding** $M(X_t)$, and **the sum of the variances of each component** $V(X_t)$. In particular, since the norm of an embedding represents the strength of its meaning (Oyama et al., 2023), $M(X_t)$ represents the strength of the meaning of the token t, while $V(X_t)$ can be interpreted as the spread of the distribution based on the variance.

In this paper, we focus on the following identity involving these three values:

$$Q(X_t) = M(X_t) + V(X_t). \tag{1}$$

As can be seen by rewriting this equation as $V(X_t) = Q(X_t) - M(X_t)$, this is nothing more than the well-known formula for variance in elementary statistics. Furthermore, we experimentally demonstrate that the variation of $Q(X_t)$ from the embeddings of intermediate layers in various Transformer models is small. Therefore, according to (1), $M(X_t)$ and $V(X_t)$ exhibit a strong trade-off relationship: when the meaning of a token is weaker, the variance of its embeddings is larger, whereas when the meaning is stronger, the variance is smaller.

To observe the trade-off relationship between $M(X_t)$ and $V(X_t)$, Fig. 1 shows PCA-transformed embeddings derived from the 6th layer of `bert-base-uncased`. We sampled tokens with frequencies evenly distributed in the range from 10^1 to 10^5 for visualization purposes (see Appendix A for more details). Tokens whose embeddings are distributed near the origin tend to have a mean embedding closer to the origin, resulting in smaller $M(X_t)$ and larger $V(X_t)$, whereas tokens whose embeddings are distributed farther from the origin have larger $M(X_t)$ and smaller $V(X_t)$. For example, the tokens *once* and *winked* have similar $Q(X_t)$ values of 494.1 and 485.6, respectively. However, the embedding set for *once* is closer to the origin than that for *winked*, with $M(X_t)$ values of 239.9 for *once* and 404.5 for *winked*. Conversely, the variance $V(X_t)$ for *once* is 254.2, larger than 81.1 for *winked*. These results are consistent with the fact that *once* functions as a stopword[2] with minimal semantic content.

To examine whether the trade-off relationship between $M(X_t)$ and $V(X_t)$, observed in Fig. 1, holds across the intermediate layers of other Transformer models, Fig. 2 presents scatter plots of

[1] Hereafter, we simply refer to "token type" as "token" for brevity.

[2] *once* is included in the stopword list provided by NLTK (Bird, 2006).

$M(X_t)$ and $V(X_t)$ for the middle-layer embeddings of six models. Consistently, the variation in $Q(X_t)$, which represents the sum of $M(X_t)$ and $V(X_t)$, remains small, confirming the trade-off relationship between $M(X_t)$ and $V(X_t)$. A detailed layer-wise analysis of this trade-off relationship is provided in Section 5.

We have obtained interesting insights not only into the set of embeddings for each token, X_t, but also into the set of embeddings for all tokens combined, $X \subset \mathbb{R}^d$. In addition to the identity similar to (1), $Q(X) = M(X) + V(X)$, we focus on the decomposition formula for variance, $V(X) = V_W(X) + V_B(X)$, where $V_W(X)$ is the within-group variance, and $V_B(X)$ is the between-group variance. Through the experiments in Section 5 using these values, we demonstrate that the embeddings in the deeper layers of Transformer models exhibit greater anisotropy.

Our main contributions are as follows:

- We focus on three statistical measures, $Q(X_t)$, $M(X_t)$, and $V(X_t)$, to analyze the distribution of contextualized embeddings. We derive the relationship in (1) and introduce an efficient method for computing $V(X_t)$ sequentially.
- We experimentally demonstrate that the variation of $Q(X_t)$ is small for embeddings from intermediate layers of various models, and that $M(X_t)$ and $V(X_t)$ exhibit a strong trade-off relationship. We theoretically argue that the Layer Normalization (LN) in BERT and RoBERTa reduces the variation of $Q(X_t)$.
- For the entire embedding set X, we derive relationships between $Q(X)$, $M(X)$, $V_W(X)$, and $V_B(X)$. We experimentally show that the layer-wise changes in these values across various Transformer models align well with previous research that highlights the anisotropy of embedding spaces.

2 Related work

2.1 Relationship between frequency and contextualized embeddings

There are three studies related to our work that deal with the relationship between word frequency and contextualized embeddings. The first is by Wannasuphoprasit et al. (2023), who found that the mean norm of BERT embeddings for the same word correlated with its frequency and proposed a frequency-considered similarity measure. In place of the mean norm, we use the mean squared norm $Q(X_t)$. The second study is by Liang et al. (2021), who demonstrated a negative correlation between word frequency and the norms of BERT embeddings. In place of the norm of the embeddings, we use the squared norm of the mean embedding $M(X_t)$. The third study is by Zhou et al. (2021, 2022a), who observed that the radius of the smallest enclosing sphere for BERT embeddings of high-frequency words tends to be larger. In place of the radius, we use the variance $V(X_t)$.

Based on these existing studies, in Section 5.4, we investigate the relationship of $Q(X_t)$, $M(X_t)$, and $V(X_t)$ against log frequency using the middle-layer embeddings of BERT.

2.2 Norms of embeddings

The norm of an embedding is an easily computed value and has been the focus of extensive research. The norm of a word embedding is related to the Kullback-Leibler divergence (Oyama et al., 2023), and embeddings of less informative words typically exhibit shorter norms (Schakel and Wilson, 2015; Arefyev et al., 2018; Kobayashi et al., 2020; Yokoi et al., 2020). Demeter et al. (2020) showed theoretically that norms are dominant in the computation of logits in the final layer. Yamagiwa et al. (2024) shows norm-derived artifacts in unnormalized embeddings, focusing on the axes of the embeddings.

2.3 Distribution of embeddings

The distribution of contextualized embeddings has been studied extensively. Contextualized embedding spaces exhibit anisotropy, primarily due to the influence of low-frequency words (Yu et al., 2022). Based on these observations, Zhang et al. (2024) proposed a method for constructing embeddings that result in an isotropic distribution. Kutuzov et al. (2022) demonstrated using ELMo (Peters et al., 2018) that embeddings of polysemous words such as *cell* form clusters according to their meanings. Yamagiwa et al. (2023) discovered that the embedding space after a whitened ICA transformation exhibits a spiky shape.

2.4 Information in layer-wise embeddings

Research focusing on the information in layer-wise embeddings is important for understanding models. Ethayarajh (2019); Cai et al. (2021); Godey et al.

(2024a) showed that the anisotropy of the embedding space increases as the layers of models such as BERT and GPT-2 deepen. Liu et al. (2019a) performed probing tasks using embeddings from different layers of ELMo, GPT-2, and BERT to investigate performance differences. Hewitt and Manning (2019) showed that the BERT embeddings from the intermediate layers capture information related to the syntax trees of sentences. Fayyaz et al. (2021) observed stability in the norms of BERT embeddings across layers. Heimersheim and Turner (2023) showed that the norm of the residual stream (Elhage et al., 2021) in GPT-2 increases as the layers deepen. Sajjad et al. (2022) showed that the variance of the embeddings differs by layer and proposed an effective post-processing.

3 Token-wise embedding set X_t

In this section, we first define the token-wise embedding set, X_t, for a given token t. Next, we provide detailed definitions of the statistical measures $Q(X_t)$, $M(X_t)$, and $V(X_t)$, and explain the relationship in (1). Finally, we show that the statistical measures of X_t can be efficiently computed through sequential computation.

3.1 Definition of X_t

We provide a formal definition of X_t, expanding on the brief explanation introduced in Section 1. Let T be the set of token types present in the corpus. For each token $t \in T$, let S_t be the set of sentences in the corpus that contain the token t. Given a contextualized embedding model f of dimension d, let $f(s,t) \in \mathbb{R}^d$ be the embedding[3] of token t in a sentence $s \in S_t$. For the token t, the set of embeddings derived from f and S_t is defined as:

$$X_t := \{f(s,t) \mid s \in S_t\} \subset \mathbb{R}^d. \quad (2)$$

We define the frequency of token t as $n_t := |X_t|$.

3.2 Statistical measures for X_t

We provide a formal definition of $Q(X_t)$, $M(X_t)$, and $V(X_t)$ for $X_t \subset \mathbb{R}^d$, and explain their relationships. First, we define the mean embedding as

$$\boldsymbol{\mu}(X_t) := \mathbb{E}_{\boldsymbol{x}\in X_t}\{\boldsymbol{x}\} = \frac{1}{|X_t|}\sum_{\boldsymbol{x}\in X_t}\boldsymbol{x} \in \mathbb{R}^d, \quad (3)$$

where $\mathbb{E}_{\boldsymbol{x}\in X_t}\{\cdot\}$ represents the sample mean over X_t. Next, for X_t, the mean squared norm $Q(X_t)$, the squared norm of the mean embedding $M(X_t)$, and the sum of the variances of each component $V(X_t)$ are defined as follows:

$$Q(X_t) := \mathbb{E}_{\boldsymbol{x}\in X_t}\left\{\|\boldsymbol{x}\|^2\right\}, \quad (4)$$

$$M(X_t) := \|\mathbb{E}_{\boldsymbol{x}\in X_t}\{\boldsymbol{x}\}\|^2 = \|\boldsymbol{\mu}(X_t)\|^2, \quad (5)$$

$$V(X_t) := \mathbb{E}_{\boldsymbol{x}\in X_t}\left\{\|\boldsymbol{x}-\boldsymbol{\mu}(X_t)\|^2\right\} = \sum_{i=1}^{d}\mathbb{E}_{\boldsymbol{x}\in X_t}\left\{(x_i-\mu_i(X_t))^2\right\}, \quad (6)$$

where x_i and $\mu_i(X_t)$ are the i-th components of $\boldsymbol{x}$ and $\boldsymbol{\mu}(X_t)$, respectively, and $\|\cdot\|$ denotes the L_2 norm. A larger $M(X_t)$ indicates that X_t is farther from the origin. Since the norm of an embedding represents the strength of its meaning (Oyama et al., 2023), a larger $M(X_t)$ indicates that token t carries greater semantic content. A larger $V(X_t)$ indicates a wider distribution within X_t, which suggests greater variability in the meaning of token t. Then, calculations (see Appendix C) yield the identity

$$Q(X_t) = M(X_t) + V(X_t),$$

which is exactly (1) in Section 1. Thus, $V(X_t)$ can be determined as $Q(X_t) - M(X_t)$, the difference between two norm-derived values.

3.3 Efficient computation for X_t

Storing all X_t when computing $Q(X_t)$, $M(X_t)$, and $V(X_t)$ is inefficient. This inefficiency can be addressed by sequentially computing $Q(X_t)$ and $\boldsymbol{\mu}(X_t)$. Using the sequentially computed $Q(X_t)$ and $\boldsymbol{\mu}(X_t)$, $M(X_t)$ and $V(X_t)$ can also be computed[4] based on (5) and (1). The procedure[5] is detailed in Algorithm 1. This algorithm requires storing only $|T|$ embeddings for $\boldsymbol{\mu}(X_t)$ and $4|T|$ scalar values, allowing for efficient handling of the statistical measures for X_t.

4 The entire embedding set X

In Section 3, we considered the embedding set X_t for each token. Considering the entire embedding set X, which includes all embedding sets X_t, we can also analyze the entire embedding space.

[3]When the same token type appears multiple times in a single sentence, embeddings are actually computed for each occurrence separately. However, for simplicity of notation, we present it as if there is a single embedding for the token in the sentence.

[4]Sequential computation methods for variance, such as Welford's online algorithm (Welford, 1962), have been known for a long time.

[5]In practice, embeddings are usually computed in batches.

Algorithm 1 Sequential computation of n_t, $\boldsymbol{\mu}(X_t)$, $Q(X_t)$, $M(X_t)$, and $V(X_t)$ for each token t

Input: A contextualized embedding model f, a corpus $S = \bigcup_{t \in T} S_t$
Output: A dictionary $\mathcal{D}$ mapping each token t to $\mathcal{D}[t] = (n_t, \boldsymbol{\mu}(X_t), Q(X_t), M(X_t), V(X_t))$

Initialize an empty dictionary $\mathcal{D}$
for each sentence $s \in S$ **do**
  for each token occurrence $t \in s$ **do**
    // Compute the token embedding
    $\boldsymbol{x} \leftarrow f(s, t) \in \mathbb{R}^d$
    if the token t is already a key in $\mathcal{D}$ **then**
      // Load previous values
      $(k, \boldsymbol{u}, q, \_, \_) \leftarrow \mathcal{D}[t]$
      // Compute new values sequentially
      $k' \leftarrow k + 1$
      $\boldsymbol{u}' \leftarrow \frac{k}{k+1}\boldsymbol{u} + \frac{1}{k+1}\boldsymbol{x} \in \mathbb{R}^d$
      $q' \leftarrow \frac{k}{k+1}q + \frac{1}{k+1}\|\boldsymbol{x}\|^2$
      $m' \leftarrow \|\boldsymbol{u}'\|^2$
      $v' \leftarrow q' - m'$
      // Update the dictionary
      $\mathcal{D}[t] \leftarrow (k', \boldsymbol{u}', q', m', v')$
    else
      // Initialize on first occurrence of t
      $\mathcal{D}[t] \leftarrow (1, \boldsymbol{x}, \|\boldsymbol{x}\|^2, \|\boldsymbol{x}\|^2, 0)$
    end if
  end for
end for

Therefore, in this section, we first provide the definition of X and then define the statistical measures for X as we did for X_t. Furthermore, we show that the total variance $V(X)$ can be decomposed into the within-group variance and the between-group variance. Finally, we explain the efficient computation for X.

4.1 Definition of X

With X_t, the entire embedding set $X \subset \mathbb{R}^d$ is defined as follows:

$$X := \bigcup_{t \in T} X_t \subset \mathbb{R}^d, \tag{7}$$

where the number of embeddings in X is defined as $n := |X| = \sum_{t \in T} n_t$.

Replacing X_t in (3) with X, we can define the mean embedding $\boldsymbol{\mu}(X) \in \mathbb{R}^d$ for X. Similarly, replacing X_t in (4), (5), and (6) with X, we can define $Q(X)$, $M(X)$, and $V(X)$, respectively. A larger $M(X)$ indicates that $\boldsymbol{\mu}(X)$ is further from the origin, making the embedding space more anisotropic. A larger $V(X)$ indicates a wider spread within the embedding space. Replacing X_t with X in (1), the following identity also holds:

$$Q(X) = M(X) + V(X). \tag{8}$$

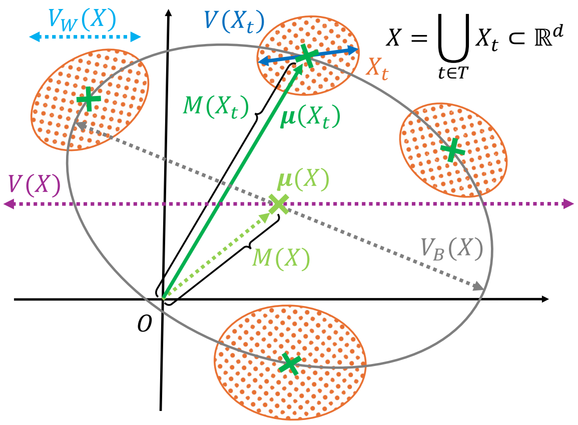


Figure 3: Illustration of the token-wise embedding sets X_t, $t \in T$, and the entire embedding set X. The values $\boldsymbol{\mu}(X_t)$, $M(X_t)$, and $V(X_t)$ are computed for each X_t, while $\boldsymbol{\mu}(X)$, $M(X)$, and $V(X)$ are for X. In addition, $V(X)$ is decomposed into the within-group variance $V_W(X)$ and the between-group variance $V_B(X)$. $V_W(X)$ is the frequency-weighted mean of $V(X_t)$, while $V_B(X)$ represents the spread of $\boldsymbol{\mu}(X_t)$ around $\boldsymbol{\mu}(X)$. Although M and V are illustrated as a norm and a standard deviation, respectively, they are actually the squared versions as shown in (5) and (6).

4.2 Within-group variance and between-group variance

In general, variance can be decomposed into within-group variance, which represents the spread within clusters, and between-group variance, which represents the spread between clusters (Muthén, 1991). Accordingly, by treating $\{X_t\}_{t \in T}$ as clusters, we consider the decomposition of the variance $V(X)$ of the entire embedding set X into the within-group variance $V_W(X)$ and the between-group variance $V_B(X)$ as follows:

$$V(X) = V_W(X) + V_B(X). \tag{9}$$

In the context of clustering, these can also be referred to as the within-cluster variance and the between-cluster variance, respectively. Calculations (see Appendix D) show that:

$$V_W(X) = \sum_{t \in T} p_t V(X_t), \tag{10}$$

$$V_B(X) = \sum_{t \in T} p_t \|\boldsymbol{\mu}(X_t) - \boldsymbol{\mu}(X)\|^2, \tag{11}$$

where

$$p_t := |X_t|/|X| = n_t/n. \tag{12}$$

Thus, $V_W(X)$ is the frequency-weighted mean of $V(X_t)$ and indicates the spread within each X_t. On the other hand, $V_B(X)$ is the frequency-weighted mean of $\|\boldsymbol{\mu}(X_t) - \boldsymbol{\mu}(X)\|^2$ and indicates the spread between $\boldsymbol{\mu}(X_t)$ around $\boldsymbol{\mu}(X)$.

From (8) and (9), the values $Q(X)$, $M(X)$, $V_W(X)$, and $V_B(X)$ satisfy:

$$Q(X) = M(X) + V_W(X) + V_B(X). \tag{13}$$

Figure 3 illustrates the relationships among these values computed from X_t and X. While the values for X_t are computed for each token, the values for X are computed from the entire embedding space.

4.3 Efficient computation for X

In Section 3.3, we showed that the statistical measures for X_t can be computed efficiently using a sequential method. Similarly, the statistical measures for X can also be computed efficiently by using the statistical measures for X_t.

As seen in Section 4.1, n can be obtained as the sum of n_t. Additionally, simple calculations (see Appendix E) show that $\boldsymbol{\mu}(X)$ and $Q(X)$ can be expressed as the frequency-weighted means of $\boldsymbol{\mu}(X_t)$ and $Q(X_t)$, respectively:

$$\boldsymbol{\mu}(X) = \mathbb{E}_{\boldsymbol{x} \in X} \{\boldsymbol{x}\} = \sum_{t \in T} p_t \boldsymbol{\mu}(X_t), \tag{14}$$

$$Q(X) = \mathbb{E}_{\boldsymbol{x} \in X} \{\|\boldsymbol{x}\|^2\} = \sum_{t \in T} p_t Q(X_t). \tag{15}$$

These expressions enable efficient computation of the statistical measures for X. Furthermore, using these values, $M(X)$, $V(X)$, $V_W(X)$, and $V_B(X)$ can also be computed efficiently.

5 Experiments

In this section, we conduct experiments using contextualized embedding models to calculate the statitical measures for X_t and X as described in Sections 3 and 4. First, we explain the experimental settings, and then present the results for X_t and X. Note that in this study, we focus on token embeddings instead of word embeddings[6], and we do not distinguish between whether a token corresponds to a complete word or a part of a word[7].

[6]This is because we found artifacts in the experimental results when representing a word embedding as the mean of the token embeddings. For details, refer to Appendix J.

[7]For example, in BERT tokenization, both the *ing* token and the *##ing* token are treated the same as the *ing* token.

Model	Layers	Dims.	Params.
`bert-base-uncased`			110M
`roberta-base`	13	768	125M
`gpt2`			117M
`bert-large-uncased`			340M
`roberta-large`	25	1024	355M
`gpt2-medium`			345M

Table 1: The number of layers including the input layer, the dimensions, and the parameter size for each model.

5.1 Settings

5.1.1 Models

We used the `transformers` library (Wolf et al., 2020) in our experiments. Following Liang et al. (2021); Zhou et al. (2022a); Wannasuphoprasit et al. (2023), we used the BERT (Devlin et al., 2019) models `bert-base-uncased` and `bert-large-uncased`. Additionally, we also used the RoBERTa (Liu et al., 2019b) models `roberta-base` and `roberta-large`, and the GPT-2 (Radford et al., 2019) models `gpt2` and `gpt2-medium`. The number of layers, the dimensions, and the size of the parameters for each model are shown in Table 1.

5.1.2 Dataset

Similar to Wannasuphoprasit et al. (2023), we used the BookCorpus (Zhu et al., 2015). For efficiency, we randomly sampled 1% of the sentences from the corpus and selected those containing fewer than 64 words for the embedding computations. The total number of sampled sentences was 739,106. Details of the number of tokens, $|T| \approx$ 24k, and the number of embeddings, $|X| \approx$ 12M, are provided in Table 3 in Appendix B. The histograms of sentence lengths and the frequency of $\log_{10} n_t$ are also shown in Figs. 9 and 10, respectively, in Appendix B.

5.2 Results for the token-wise embedding sets

Figure 2 shows scatter plots of $V(X_t)$ against $M(X_t)$ from the middle-layer embeddings of the six models. Each scatter plot shows the regression line and displays its slope and the coefficient of determination, R^2. Consistently, the sum of $M(X_t)$ and $V(X_t)$, namely $Q(X_t)$, exhibits small variation, confirming the trade-off relationship between $M(X_t)$ and $V(X_t)$. Furthermore, the slopes of the regression lines are negative, with large R^2 val-

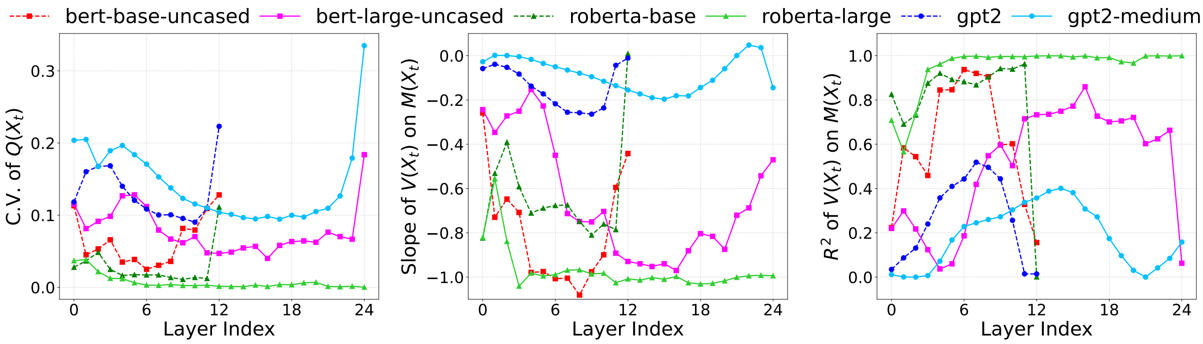


Figure 4: For each layer across the six models, the coefficient of variation (C.V.) of $Q(X_t)$ on the left, the slope of the regression line of $V(X_t)$ on $M(X_t)$ in the middle, and the corresponding coefficient of determination R^2 on the right are shown. For all models, the C.V. approximately reaches its minimum in the intermediate layers. Consequently, the slope and R^2 approximately reach their minimum and maximum, respectively, in the intermediate layers. Only tokens with $1 \leq \log_{10} n_t \leq 5$ were used to reduce the influence of extreme values.

ues. For example, in the case of `roberta-large`, the slope of the regression line is -1.008 and $R^2 = 0.999$, indicating a nearly perfect trade-off relationship with a constant sum.

Next, we examine the variation of $Q(X_t)$ and the trade-off between $M(X_t)$ and $V(X_t)$ across layers. Figure 4 shows the coefficient of variation[8] (C.V.) of $Q(X_t)$, the slope of the regression line of $V(X_t)$ on $M(X_t)$, and the corresponding R^2 value for each layer of the six models. The C.V. of $Q(X_t)$ is generally low and it reaches its minimum value approximately in the intermediate layers of each model, where the trade-off between $M(X_t)$ and $V(X_t)$ becomes more pronounced. In the intermediate layers of BERT and RoBERTa, the slope of the regression line reaches a minimum value of approximately -1, and the R^2 value approaches its maximum of 1. However, in the case of GPT-2, the minimum C.V. of $Q(X_t)$ is larger than those of BERT and RoBERTa, with a minimum slope of approximately -0.2 and a maximum R^2 value of around 0.5. These differences are likely due to architectural differences, which will be discussed in Section 6.

5.3 Results for the entire embedding set

As seen in (13), $Q(X)$ can be decomposed into $M(X)$, $V_W(X)$, and $V_B(X)$. Figure 5 illustrates the changes in the ratios of $M(X)$, $V_W(X)$, and $V_B(X)$ normalized by $Q(X)$ across the layers of the six models. Generally, as the layers deepen, the ratio of $M(X)$ increases, which means that the ratio of the sum $V_W(X) + V_B(X)$ (equal to $V(X)$) decreases. Additionally, a comparison between $V_W(X)$ and $V_B(X)$ shows that the ratio of $V_W(X)$ increases as the layers deepen. Figure 19 in Appendix G presents the original layer-wise values of $Q(X)$, $M(X)$, and $V(X)$.

According to (9), $V_W(X) + V_B(X) = V(X)$. To investigate the value of $V_W(X)$ relative to $V_B(X)$, Fig. 6 shows the ratio $V_W(X)/V(X)$. Consistent with the results in Fig. 5, the ratio of $V_W(X)$ increases in each model as the layers deepen, i.e., the ratio of $V_B(X)$ decreases.

Previous studies on the anisotropy of embedding spaces across layers (Ethayarajh, 2019; Cai et al., 2021; Godey et al., 2024a) showed that for BERT, RoBERTa, and GPT-2, the average cosine similarity between randomly sampled words increases as the layers deepen. This finding is consistent with our results in Fig. 5, where the ratio of $M(X)$ increases and the ratio of $V(X)$ decreases as the layers deepen, and in Fig. 6, where the ratio of $V_B(X)$ decreases. These studies also found that the cosine similarity between embeddings of the same word in different sentences decreases as the layers deepen. This observation is also consistent with our results in Fig. 6, where the ratio of $V_W(X)$ gradually increases. While previous work such as Ethayarajh (2019) computed cosine similarities by randomly sampling 1,000 embeddings, we computed the values using all embeddings in the dataset.

5.4 Relationship of $Q(X_t)$, $M(X_t)$, and $V(X_t)$ against token frequency

In Section 2.1, we discussed three related studies that examined the relationship between word frequency and values associated with $Q(X_t)$, $M(X_t)$,

[8] C.V. is defined as the ratio of the standard deviation to the mean, representing the relative variability in the data.

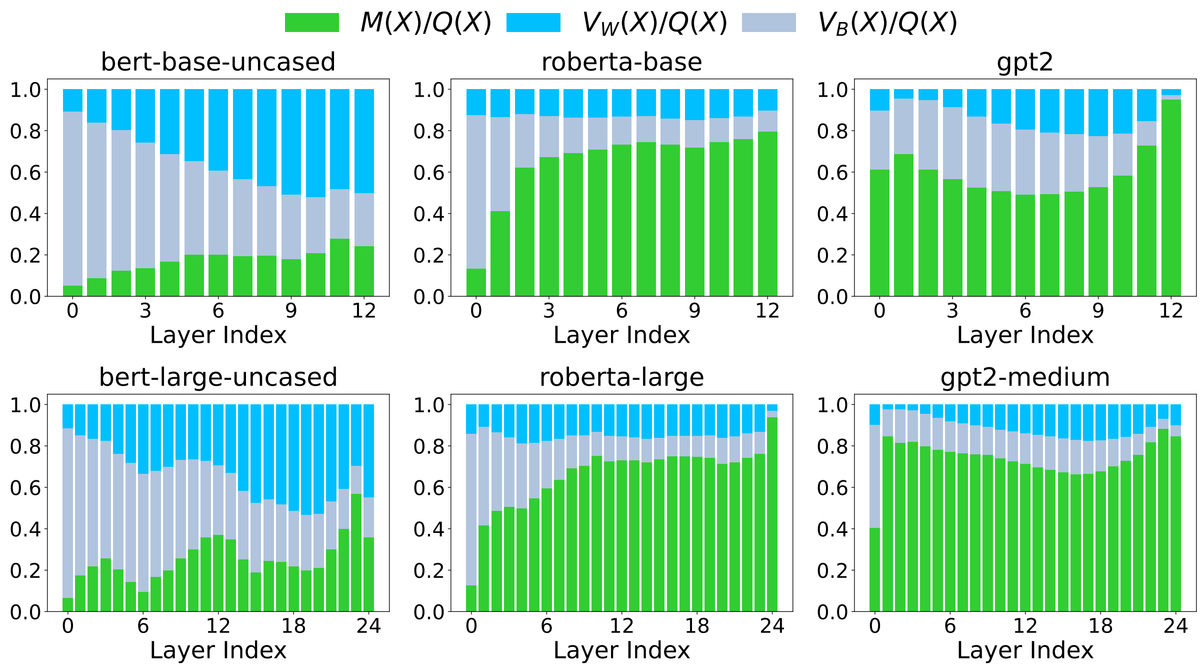


Figure 5: The ratios of $M(X)$, $V_W(X)$, and $V_B(X)$, each normalized by $Q(X)$, for each layer across the six models. As the layers deepen, the ratio of $M(X)$ tends to exceed that of $V_W(X) + V_B(X)(= V(X))$. Meanwhile, the ratio of $V_W(X)$ increases relative to $V_B(X)$. Figure 6 shows detailed comparisons between $V_W(X)$ and $V_B(X)$. Further plots of the ratios of these values and those of the original values are shown in Figs. 18 and 19, respectively, in Appendix G. Only tokens with $1 \leq \log_{10} n_t \leq 5$ were used to reduce the influence of extreme values.

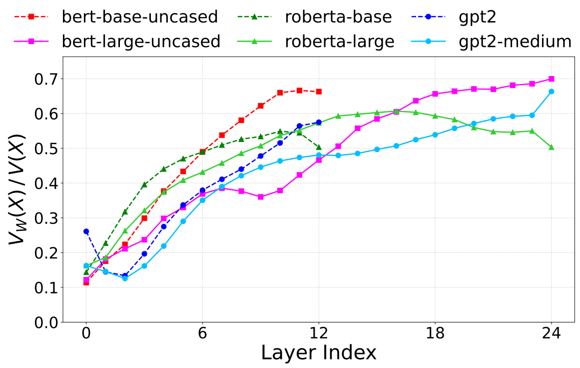


Figure 6: The ratio $V_W(X)/V(X)$ in Fig. 5 for each layer across the six models. As the layers deepen, the ratio of $V_W(X)$ increases. Further plots of these values are shown in Fig. 20 in Appendix G.

and $V(X_t)$. In this section, we examine the correlations between these three proposed values and token frequency.

Figure 7 presents scatter plots of $Q(X_t)$, $M(X_t)$, and $V(X_t)$ against log frequency, using embeddings from the 6th layer of `bert-base-uncased` as a representative example. The slope of $Q(X_t)$ remains stable and close to zero. In contrast, the negative slope of $M(X_t)$ and the positive slope of $V(X_t)$ indirectly suggest a trade-off relationship between $M(X_t)$ and $V(X_t)$. Similar trends were observed across different layers and models (see Appendix F).

6 Discussion

6.1 Why does the C.V. of $Q(X_t)$ reach its minimum in the intermediate layers?

In this study, we discovered that the C.V. of $Q(X_t)$ decreases in the intermediate layers of various Transformer models, as shown in Fig. 4. While further investigation into the reasons behind this phenomenon remains as future work, we present our hypothesis here.

In the input layer (layer 0), pre-trained embeddings are used, whereas in the final layer, embeddings are influenced by the objective function and the computation of logits. As a result, embeddings in these two layers are expected to exhibit different characteristics compared to those in other layers. Indeed, as shown in Fig. 19 in Appendix G, the plots of $Q(X)$, $M(X)$, and $V(X)$ for each layer indicate that the values in the input and output layers differ significantly from those in other layers. This suggests that the influence of these specialized layers is reduced in the intermediate layers, possibly reflecting a property inherent to the model architecture or language — namely, the reduced

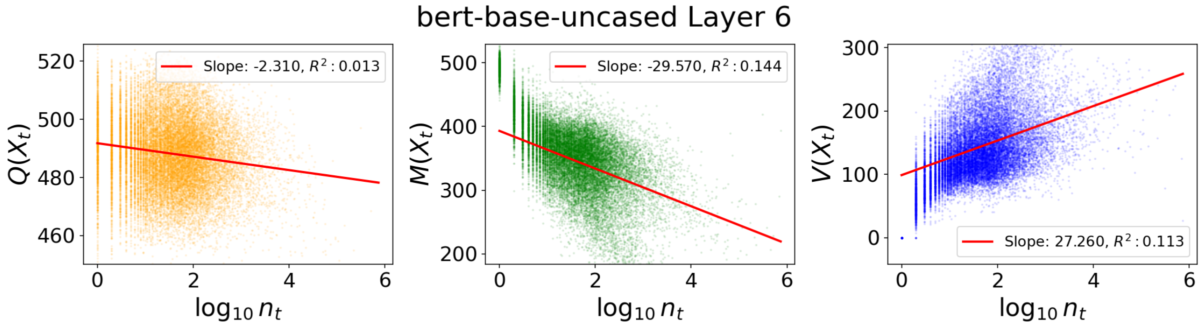


Figure 7: Scatter plots of (left) $Q(X_t)$, (middle) $M(X_t)$, and (right) $V(X_t)$ against $\log_{10} n_t$ for the middle-layer embeddings of `bert-base-uncased`. Each plot includes a regression line, its slope, and the coefficient of determination (R^2). While the slope of $Q(X_t)$ is close to zero, those of $M(X_t)$ and $V(X_t)$ are negative and positive, respectively. Only tokens with $1 \leq \log_{10} n_t \leq 5$ were used for regressions to reduce the influence of extreme values. Appendix F presents the results for embeddings from multiple layers of several models based on log-scaled values.

variation in $Q(X_t)$.

6.2 Why does GPT-2 behave differently from BERT and RoBERTa in Fig. 4?

In Fig. 4, although GPT-2 shows similar trends to BERT and RoBERTa, the minimum C.V. of $Q(X_t)$, the minimum slope of the regression line, and the maximum R^2 differ from those of BERT and RoBERTa. Furthermore, for GPT-2, the values of $Q(X_t)$, $M(X_t)$, and $V(X_t)$ in Fig. 2, as well as $Q(X)$, $M(X)$, and $V(X)$ in Fig. 19 in Appendix G, differ significantly in magnitude from those for BERT and RoBERTa.

This is likely due to the different placement of layer normalization (LN) within the Transformer layer. In general, as shown in Fig. 21 in Appendix H, Post-LN Transformers such as BERT and RoBERTa apply the LN after the feed-forward network (FF), while Pre-LN Transformers such as GPT-2 apply it before the FF (Xiong et al., 2020). Consequently, the embeddings $\boldsymbol{x} \in X_t$ of BERT and RoBERTa are outputs of the LN, and the squared norm $\|\boldsymbol{x}\|^2$ is controlled with small variation.

For Post-LN Transformers, under certain specific conditions assuming an ideal scenario, it can be shown that the C.V. of $Q(X_t)$ is sufficiently small. As shown in Appendix I,

$$C.V.(Q(X_t)) = O\left(\frac{1}{\sqrt{n_0}}\right), \qquad (16)$$

where $n_0 = \min_{t \in T} n_t$. This result indicates that the C.V. approaches zero as all n_t increase.

On the other hand, $\|\boldsymbol{x}\|^2$ for GPT-2 is not controlled by the LN, yet it is interesting to observe similar trends in Figs. 5 and 6. Although it is not necessarily desirable for embeddings to be artificially constrained directly by LN, as in BERT and RoBERTa, the trade-off between $M(X_t)$ and $V(X_t)$ is also observed in the GPT-2 model, which is not directly constrained in this way. This observation suggests that the constraints imposed by LN reflect reality to some extent and did not cause a significant issue in the model's language learning process.

7 Conclusion

In this study, we focused on the distribution of contextualized embeddings and analyzed three values: the mean squared norm $Q(X_t)$, the squared norm of the mean embedding $M(X_t)$, and the sum of the variances of each component $V(X_t)$. In Section 3, we showed that the values of $Q(X_t)$, $M(X_t)$, and $V(X_t)$ are related by (1) and can be efficiently computed using a sequential method. We also found that, in the intermediate layers of several models, the variation of $Q(X_t)$ is small, which results in a strong trade-off between $M(X_t)$ and $V(X_t)$. We explained in Section 6 that the small variation in $Q(X_t)$ can be attributed to the placement of LN. The values of $Q(X_t)$, $M(X_t)$, and $V(X_t)$ can also be applied to the entire embedding set X, and we demonstrated that the total variance $V(X)$ can be decomposed into within-group variance $V_W(X)$ and between-group variance $V_B(X)$. As seen in Figs. 5 and 6, the experimental results from relative comparisons show that as the layers deepen, $M(X)$ increases, while $V(X)$ and $V_B(X)$ decrease, and $V_W(X)$ increases. These results are consistent with existing studies on the anisotropy of embedding spaces across layers.

Limitations

- Due to computational resource limitations, we used relatively small models with parameter sizes fewer than 1B, as shown in Table 1. Since anisotropy in embedding spaces is affected by parameter size (Godey et al., 2024b), verification with larger models would be desirable. Note that previous studies on the relationship between word frequency and embeddings (Liang et al., 2021; Zhou et al., 2022a; Wannasuphoprasit et al., 2023) have only examined the final layer of BERT models. In contrast, we conducted experiments with BERT, RoBERTa, and GPT-2, following the settings from previous work on layer-wise anisotropy (Ethayarajh, 2019; Cai et al., 2021; Godey et al., 2024a).

- In order to run the experiments efficiently, we did not use the full BookCorpus. The number of sentences used in the experiments was 739,106, and the total number of embeddings $|X|$ exceeded 10 million, which we considered sufficient.

- This study deals only with English models. The analysis of the values of Q, M, and V for different languages using multilingual models is left for future work.

- In this study, we analyzed high-dimensional distributions using scalar values such as norms and the sum of variances, prioritizing ease of interpretation, as discussed in Sections 3 and 4.

- Ethayarajh (2019); Cai et al. (2021); Godey et al. (2024a) used cosine similarity to examine layer-wise anisotropy, which facilitates comparisons across models and layers. In contrast, the values of Q, M, and V are not normalized, and these values vary significantly across models and layers. Therefore, appropriate adjustments may be necessary for such comparisons. Based on this, in Figs. 5 and 6, we have normalized the values using $Q(X)$ and $V(X)$ to allow comparisons across models and layers. This normalization shows, for example, that while the value of $V(X)$ increases as the layers deepen in GPT-2 (Fig. 19 in Appendix G), the ratio of $V(X)/Q(X)$ decreases, as shown in Fig. 5.

- In our experiments, only tokens with $1 \leq \log_{10} n_t \leq 5$ were used when computing values such as the regression line, to reduce the influence of extreme values. The choice of this frequency range is ad hoc, and the influence of token frequency on the results has not been examined in detail.

- The probability distribution settings assumed in the theory of the C.V. of $Q(X_t)$ (Appendix I) do not necessarily reflect reality, and the derived formulas have only limited value.

- A more detailed experimental and theoretical analysis is needed to understand why the C.V. of $Q(X_t)$ becomes smaller in the intermediate layers and why GPT-2 behaves differently from BERT and RoBERTa in Fig. 4.

Ethics Statement

This study complies with the ACL Ethics Policy.

Acknowledgments

We would like to thank Yusuke Takase, Momose Oyama, Ryo Kishino, and the anonymous reviewers for their helpful comments and suggestions. This study was partially supported by JSPS KAKENHI 22H05106, 23H03355, JST CREST JPMJCR21N3, JST SPRING JPMJSP2110.

References

Nikolay Arefyev, Pavel Ermolaev, and Alexander Panchenko. 2018. How much does a word weigh? weighting word embeddings for word sense induction. *CoRR*, abs/1805.09209.

Steven Bird. 2006. NLTK: the natural language toolkit. In *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*. The Association for Computer Linguistics.

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. 2021. Isotropy in the contextual embedding space: Clusters and manifolds. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

David Demeter, Gregory Kimmel, and Doug Downey. 2020. Stolen probability: A structural weakness of neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2191–2197. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 55–65. Association for Computational Linguistics.

Mohsen Fayyaz, Ehsan Aghazadeh, Ali Modarressi, Hosein Mohebbi, and Mohammad Taher Pilehvar. 2021. Not all models localize linguistic knowledge in the same place: A layer-wise probing on bertoids' representations. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2021, Punta Cana, Dominican Republic, November 11, 2021*, pages 375–388. Association for Computational Linguistics.

Nathan Godey, Éric Villemonte de la Clergerie, and Benoît Sagot. 2024a. Anisotropy is inherent to self-attention in transformers. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers, St. Julian's, Malta, March 17-22, 2024*, pages 35–48. Association for Computational Linguistics.

Nathan Godey, Éric Villemonte de la Clergerie, and Benoît Sagot. 2024b. Why do small language models underperform? studying language model saturation via the softmax bottleneck. In *First Conference on Language Modeling*.

Stefan Heimersheim and Alex Turner. 2023. Residual stream norms grow exponentially over the forward pass. *LESSWRONG*.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4129–4138. Association for Computational Linguistics.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7057–7075. Association for Computational Linguistics.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2024. Analyzing feed-forward blocks in transformers through the lens of attention maps. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. 2022. Contextualized embeddings for semantic change detection: Lessons learned. In *Northern European Journal of Language Technology, Volume 8*, Copenhagen, Denmark. Northern European Association of Language Technology.

Yuxin Liang, Rui Cao, Jie Zheng, Jie Ren, and Ling Gao. 2021. Learning to remove: Towards isotropic pre-trained BERT embedding. In *Artificial Neural Networks and Machine Learning - ICANN 2021 - 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14-17, 2021, Proceedings, Part V*, volume 12895 of *Lecture Notes in Computer Science*, pages 448–459. Springer.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1073–1094. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.

Bengt O Muthén. 1991. Multilevel factor analysis of class and student achievement components. *Journal of Educational measurement*, 28(4):338–354.

Momose Oyama, Sho Yokoi, and Hidetoshi Shimodaira. 2023. Norm of word embedding encodes information gain. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2108–2130. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Hassan Sajjad, Firoj Alam, Fahim Dalvi, and Nadir Durrani. 2022. Effect of post-processing on contextualized word representations. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 3127–3142. International Committee on Computational Linguistics.

Adriaan M. J. Schakel and Benjamin J. Wilson. 2015. Measuring word significance using distributed representations of words. *CoRR*, abs/1508.02297.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *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.

Saeth Wannasuphoprasit, Yi Zhou, and Danushka Bollegala. 2023. Solving cosine similarity underestimation between high frequency words by ℓ_2 norm discounting. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8644–8652. Association for Computational Linguistics.

Barry Payne Welford. 1962. Note on a method for calculating corrected sums of squares and products. *Technometrics*, 4(3):419–420.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. 2020. On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 10524–10533. PMLR.

Hiroaki Yamagiwa, Momose Oyama, and Hidetoshi Shimodaira. 2023. Discovering universal geometry in embeddings with ICA. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 4647–4675. Association for Computational Linguistics.

Hiroaki Yamagiwa, Momose Oyama, and Hidetoshi Shimodaira. 2024. Revisiting cosine similarity via normalized ica-transformed embeddings. *CoRR*, abs/2406.10984.

Sho Yokoi, Ryo Takahashi, Reina Akama, Jun Suzuki, and Kentaro Inui. 2020. Word rotator's distance. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2944–2960. Association for Computational Linguistics.

Sangwon Yu, Jongyoon Song, Heeseung Kim, Seongmin Lee, Woo-Jong Ryu, and Sungroh Yoon. 2022. Rare tokens degenerate all tokens: Improving neural text generation via adaptive gradient gating for rare token embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 29–45. Association for Computational Linguistics.

Ying Zhang, Dongyuan Li, and Manabu Okumura. 2024. Reconsidering token embeddings with the definitions for pre-trained language models. *Preprint*, arXiv:2408.01308.

Kaitlyn Zhou, Kawin Ethayarajh, Dallas Card, and Dan Jurafsky. 2022a. Problems with cosine as a measure of embedding similarity for high frequency words. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 401–423. Association for Computational Linguistics.

Kaitlyn Zhou, Kawin Ethayarajh, and Dan Jurafsky. 2021. Frequency-based distortions in contextualized word embeddings. *CoRR*, abs/2104.08465.

Kaitlyn Zhou, Kawin Ethayarajh, and Dan Jurafsky. 2022b. Richer countries and richer representations.

In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2074–2085. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 19–27. IEEE Computer Society.

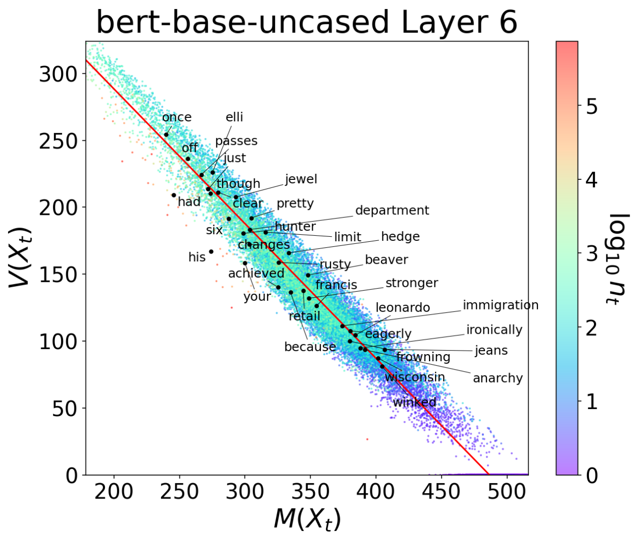


Figure 8: Scatter plots of $M(X_t)$ and $V(X_t)$ using the 6th layer of `bert-base-uncased`, along with the regression line and selected tokens from Fig. 1. The tokens are close to the regression line, indicating that the tokens have similar $Q(X_t)$ values.

token t	n_t	$Q(X_t)$	$M(X_t)$	$V(X_t)$
his	94718	441.0	273.9	167.1
had	56623	454.8	245.7	209.1
just	22619	485.4	271.9	213.5
your	20307	458.2	299.9	158.3
off	12005	492.7	256.3	236.4
because	8484	471.2	334.7	136.5
though	5085	484.1	273.7	210.4
once	5022	494.1	239.9	254.2
pretty	2082	496.7	304.7	192.0
clear	1479	490.7	279.4	211.3
six	1302	479.2	287.6	191.6
hunter	770	478.9	298.5	180.4
jeans	727	500.3	406.6	93.7
stronger	435	480.6	348.6	132.0
department	233	487.0	303.7	183.3
winked	229	485.6	404.5	81.1
changes	214	475.7	303.4	172.3
frowning	210	479.8	379.8	100.0
jewel	157	500.2	292.7	207.5
rusty	124	484.3	325.5	158.8
eagerly	122	488.0	380.4	107.6
passes	118	490.9	266.6	224.3
limit	88	496.9	315.5	181.4
elli	61	501.3	275.2	226.1
hedge	59	499.4	333.5	165.9
francis	57	480.8	354.5	126.3
achieved	35	465.4	325.1	140.3
ironically	33	483.0	388.2	94.8
beaver	26	497.3	348.0	149.3
leonardo	23	488.5	384.1	104.4
immigration	17	485.3	374.0	111.3
anarchy	12	484.7	391.3	93.4
retail	10	482.1	344.5	137.6
wisconsin	10	488.7	401.6	87.1

Table 2: Values of n_t, $Q(X_t)$, $M(X_t)$, and $V(X_t)$ for the tokens in Fig. 1. See Appendix A for details on the token selection method.

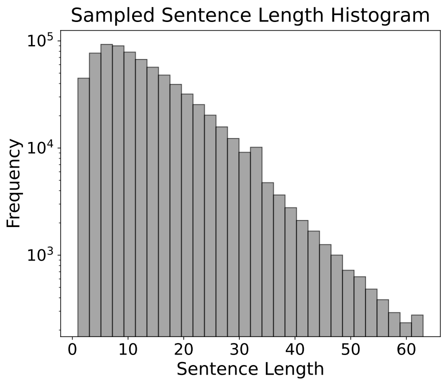


Figure 9: Histograms of sentence lengths for the dataset we used. See Appendix B for details on how the sentences were selected.

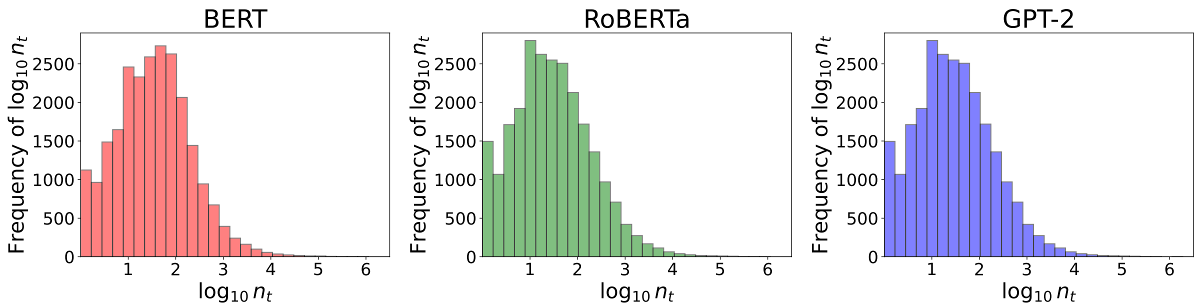


Figure 10: Histograms of $\log_{10} n_t$ for each model. For differences in tokenization for each model, see Appendix B.

A Details of Fig. 1

In this section, we explain the selection process of the tokens used in Fig. 1. To account for the effect of frequency, we used only tokens with $1 \leq \log_{10} n_t \leq 5$, and for readability, we limited the selection to tokens with at least 3 characters. Based on these conditions, we defined the interval $[\min_t \log_{10} n_t, \max_t \log_{10} n_t]$ and divided it into 10 equal subintervals. Let I_r denote the r-th subinterval, where $r \in \{1, \dots, 10\}$. From each I_r, we defined the set of tokens as:

$$T_r := \{t \in T \mid \log_{10}(n_t) \in I_r\}.$$

Let $|T_r|$ denote the number of tokens in T_r. For each T_r, we ramdomly sampled

$$N_r := 2 + \left\lfloor \sqrt{\frac{4\,|T_r|}{\max_r |T_r|}} \right\rfloor$$

tokens. Here, $\lfloor \cdot \rfloor$ represents the floor function. The definition of N_r is ad hoc, ensuring that at least two tokens are sampled from each T_r, with additional tokens sampled proportionally to $|T_r|$.

In Fig. 8, the selected tokens (i.e., the tokens shown in Fig. 1) are plotted in scatter plots of $M(X_t)$ and $V(X_t)$.

Table 2 shows the values of n_t, $Q(X_t)$, $M(X_t)$, and $V(X_t)$ for these tokens. Despite large differences in frequency, these tokens have similar values for $Q(X_t)$.

For the PCA transformation used in Fig. 1, we also transformed the origin $\mathbf{0} \in \mathbb{R}^d$ to better understand the positional relationship between each distribution of embeddings and the origin. After the transformation, we translated all 2D points so that the transformed origin coincided with the new origin.

B Details of the dataset

As described in Section 5.1.2, we randomly sampled 1% of the sentences from the BookCorpus (Zhu et al., 2015) and used sentences with fewer than 64 words for embedding calculations. The histogram of sentence lengths for the sampled sentences is shown in Fig. 9.

Model	$\|T\|$	$\|X\|$
BERT	24,149	12,384,011
RoBERTa	24,719	12,238,028
GPT-2	24,718	12,238,028

Table 3: Values of $|T|$ and $|X|$ by tokenization for each model. While RoBERTa uses the same tokenizer as GPT-2, it distinguishes between the beginning-of-sentence and end-of-sentence tokens, unlike GPT-2.

The tokenization of BERT differs from that of RoBERTa and GPT-2. RoBERTa uses the same tokenizer as GPT-2 but differs in that it distinguishes between beginning-of-sentence (BOS) and end-of-sentence (EOS) tokens. RoBERTa uses *<s>* for BOS and *</s>* for EOS, while GPT-2 uses *<|endoftext|>* for both BOS and EOS. Table 3 shows $|T|$ and $|X|$. The histograms of log-scale token frequencies, $\log_{10} n_t$, for each model are shown in Fig. 10.

C Details of the statistical measures for an embedding set X_t in Section 3

In this section, as discussed in Section 3, we explain the values $Q(X_t)$, $M(X_t)$, and $V(X_t)$ for an embedding set X_t, as well as the relationship between them, given by (1). First, we explain the statistical measures for each component, and then we provide the proof of (1).

C.1 Values for each component

Using the i-th component x_i of an embedding $\boldsymbol{x} \in \mathbb{R}^d$, we define the following values for X_t:

$$q_i(X_t) := \mathbb{E}_{\boldsymbol{x}\in X_t}\left\{x_i^2\right\}, \tag{17}$$

$$\mu_i(X_t) := \mathbb{E}_{\boldsymbol{x}\in X_t}\left\{x_i\right\}, \tag{18}$$

$$v_i(X_t) := \mathbb{E}_{\boldsymbol{x}\in X_t}\left\{(x_i - \mu_i(X_t))^2\right\}, \tag{19}$$

where $\mu_i(X_t)$ is the i-th component of $\boldsymbol{\mu}(X_t)$ in (3), and $v_i(X_t)$ is the variance of the i-th component of the embeddings in X_t, $\{x_i \mid \boldsymbol{x} \in X_t\}$. Then, the following relationship holds between $q_i(X_t)$, $\mu_i(X_t)$, and $v_i(X_t)$:

$$q_i(X_t) = \mu_i(X_t)^2 + v_i(X_t). \tag{20}$$

Proof.

$$\begin{aligned} v_i(X_t) &= \mathbb{E}_{\boldsymbol{x}\in X_t}\left\{(x_i - \mu_i(X_t))^2\right\} \\ &= \mathbb{E}_{\boldsymbol{x}\in X_t}\left\{x_i^2 - 2\mu_i(X_t)x_i + \mu_i(X_t)^2\right\} \\ &= \mathbb{E}_{\boldsymbol{x}\in X_t}\left\{x_i^2\right\} - \mu_i(X_t)^2 \\ &= q_i(X_t) - \mu_i(X_t)^2. \end{aligned}$$

□

This is nothing more than the well-known formula for variance in elementary statistics.

Thus, $Q(X_t)$, $M(X_t)$, and $V(X_t)$ are the sums of $q_i(X_t)$, $\mu_i(X_t)^2$, and $v_i(X_t)$ across all components, as follows:

$$Q(X_t) = \sum_{i=1}^{d} q_i(X_t), \tag{21}$$

$$M(X_t) = \sum_{i=1}^{d} \mu_i(X_t)^2, \tag{22}$$

$$V(X_t) = \sum_{i=1}^{d} v_i(X_t). \tag{23}$$

Proof. From the definition of the L_2 norm:

$$\|\boldsymbol{x}\|^2 = \sum_{i=1}^{d} x_i^2.$$

Then we obtain:

$$\begin{aligned}
Q(X_t) &= \mathbb{E}_{\boldsymbol{x}\in X_t}\left\{\|\boldsymbol{x}\|^2\right\} \\
&= \mathbb{E}_{\boldsymbol{x}\in X_t}\left\{\sum_{i=1}^{d} x_i^2\right\} \\
&= \sum_{i=1}^{d} \mathbb{E}_{\boldsymbol{x}\in X_t}\left\{x_i^2\right\} \\
&= \sum_{i=1}^{d} q_i(X_t), \\
M(X_t) &= \|\mathbb{E}_{\boldsymbol{x}\in X_t}\left\{\boldsymbol{x}\right\}\|^2 \\
&= \|\boldsymbol{\mu}(X_t)\|^2 \\
&= \sum_{i=1}^{d} \mu_i(X_t)^2, \\
V(X_t) &= \mathbb{E}_{\boldsymbol{x}\in X_t}\left\{\|\boldsymbol{x} - \boldsymbol{\mu}(X_t)\|^2\right\} \\
&= \mathbb{E}_{\boldsymbol{x}\in X_t}\left\{\sum_{i=1}^{d} (x_i - \mu_i(X_t))^2\right\} \\
&= \sum_{i=1}^{d} \mathbb{E}_{\boldsymbol{x}\in X_t}\left\{(x_i - \mu_i(X_t))^2\right\} \\
&= \sum_{i=1}^{d} v_i(X_t).
\end{aligned}$$

□

C.2 Proof of (1) related to $Q(X_t)$, $M(X_t)$, and $V(X_t)$

We prove (1) given by:

$$Q(X_t) = M(X_t) + V(X_t).$$

Proof. By summing both sides of (20) from $i = 1$ to d:

$$\sum_{i=1}^{d} q_i(X_t) = \sum_{i=1}^{d} \mu_i(X_t)^2 + \sum_{i=1}^{d} v_i(X_t).$$

Then we obtain the result using (21), (22), and (23). Alternatively, the result can be derived directly as follows:

$$
\begin{aligned}
V(X_t) &= \mathbb{E}_{\boldsymbol{x} \in X_t} \left\{ \|\boldsymbol{x} - \boldsymbol{\mu}(X_t)\|^2 \right\} \\
&= \mathbb{E}_{\boldsymbol{x} \in X_t} \left\{ (\boldsymbol{x} - \boldsymbol{\mu}(X_t))^\top (\boldsymbol{x} - \boldsymbol{\mu}(X_t)) \right\} \\
&= \mathbb{E}_{\boldsymbol{x} \in X_t} \left\{ \|\boldsymbol{x}\|^2 - \boldsymbol{x}^\top \boldsymbol{\mu}(X_t) - \boldsymbol{\mu}(X_t)^\top \boldsymbol{x} + \|\boldsymbol{\mu}(X_t)\|^2 \right\} \\
&= \mathbb{E}_{\boldsymbol{x} \in X_t} \left\{ \|\boldsymbol{x}\|^2 \right\} - \mathbb{E}_{\boldsymbol{x} \in X_t} \left\{ \boldsymbol{x} \right\}^\top \boldsymbol{\mu}(X_t) - \boldsymbol{\mu}(X_t)^\top \mathbb{E}_{\boldsymbol{x} \in X_t} \left\{ \boldsymbol{x} \right\} + \|\boldsymbol{\mu}(X_t)\|^2 \\
&= \mathbb{E}_{\boldsymbol{x} \in X_t} \left\{ \|\boldsymbol{x}\|^2 \right\} - \|\boldsymbol{\mu}(X_t)\|^2 \\
&= Q(X_t) - M(X_t).
\end{aligned}
$$

□

D Proof of (9) decomposing $V(X)$ into $V_W(X)$ and $V_B(X)$

In this section, we will prove (9). To do so, we first prove the following equation for $v_i(X)$, where $v_i(X)$ is the value obtained by replacing X_t with X in $v_i(X_t)$ as defined in (19) in Appendix C:

$$
v_i(X) = \sum_{t \in T} p_t \left\{ v_i(X_t) + (\mu_i(X_t) - \mu_i(X))^2 \right\}. \tag{24}
$$

Proof.

$$
\begin{aligned}
v_i(X) &= \mathbb{E}_{\boldsymbol{x} \in X} \left\{ (x_i - \mu_i(X))^2 \right\} \\
&= \frac{1}{|X|} \sum_{\boldsymbol{x} \in X} (x_i - \mu_i(X))^2 \\
&= \frac{1}{|X|} \sum_{t \in T} \sum_{\boldsymbol{x} \in X_t} (x_i - \mu_i(X))^2 \\
&= \sum_{t \in T} \frac{|X_t|}{|X|} \cdot \frac{1}{|X_t|} \sum_{\boldsymbol{x} \in X_t} (x_i - \mu_i(X))^2 \\
&= \sum_{t \in T} p_t \mathbb{E}_{\boldsymbol{x} \in X_t} \left\{ (x_i - \mu_i(X_t) + \mu_i(X_t) - \mu_i(X))^2 \right\} \\
&= \sum_{t \in T} p_t \mathbb{E}_{\boldsymbol{x} \in X_t} \left\{ (x_i - \mu_i(X_t))^2 - 2(x_i - \mu_i(X_t))(\mu_i(X_t) - \mu_i(X)) + (\mu_i(X_t) - \mu_i(X))^2 \right\} \\
&= \sum_{t \in T} p_t \left\{ \mathbb{E}_{\boldsymbol{x} \in X_t} \left\{ (x_i - \mu_i(X_t))^2 \right\} + (\mu_i(X_t) - \mu_i(X))^2 \right\} \quad (\because \mu_i(X_t) = \mathbb{E}_{\boldsymbol{x} \in X_t} \left\{ x_i \right\}) \\
&= \sum_{t \in T} p_t \left\{ v_i(X_t) + (\mu_i(X_t) - \mu_i(X))^2 \right\}.
\end{aligned}
$$

□

Based on (24), we prove (9) given by:

$$
V(X) = V_W(X) + V_B(X).
$$

Proof.

$$\begin{aligned} V(X) &= \sum_{i=1}^{d} v_i(X) \quad \text{(from (23), where } X_t \text{ is replaced by } X) \\ &= \sum_{i=1}^{d} \sum_{t \in T} p_t \left\{ v_i(X_t) + (\mu_i(X_t) - \mu_i(X))^2 \right\} \\ &= \sum_{t \in T} p_t \sum_{i=1}^{d} v_i(X_t) + \sum_{t \in T} p_t \sum_{i=1}^{d} (\mu_i(X_t) - \mu_i(X))^2 \\ &= \sum_{t \in T} p_t V(X_t) + \sum_{t \in T} p_t \|\boldsymbol{\mu}(X_t) - \boldsymbol{\mu}(X)\|^2 \quad (\because (23)) \\ &= V_W(X) + V_B(X). \end{aligned}$$

□

E Details of the statistical measures for the entire embedding set X in Section 4.3

In this section, as discussed in Section 4.3, we explain the values n, $\boldsymbol{\mu}(X)$, and $Q(X)$ for the entire embedding set X.

E.1 Calculation of n

We define $n = |X|$ as the total number of embeddings in X. We assume that X_t and $X_{t'}$ are disjoint for $t, t' \in T$, i.e., $X_t \cap X_{t'} = \emptyset$. Then

$$n = \left| \bigcup_{t \in T} X_t \right| = \sum_{t \in T} |X_t| = \sum_{t \in T} n_t.$$

Thus, n can be expressed using n_t.

E.2 Calculation of $\boldsymbol{\mu}(X)$

By replacing X_t with X in $\boldsymbol{\mu}(X_t)$ in (3), the mean embedding $\boldsymbol{\mu}(X) \in \mathbb{R}^d$ for X is defined. We prove (14) given by:

$$\boldsymbol{\mu}(X) = \mathbb{E}_{\boldsymbol{x} \in X} \left\{ \boldsymbol{x} \right\} = \sum_{t \in T} p_t \boldsymbol{\mu}(X_t).$$

Proof.

$$\begin{aligned} \boldsymbol{\mu}(X) &= \mathbb{E}_{\boldsymbol{x} \in X} \{\boldsymbol{x}\} \\ &= \frac{1}{|X|} \sum_{\boldsymbol{x} \in X} \boldsymbol{x} \\ &= \frac{1}{|X|} \sum_{t \in T} \sum_{\boldsymbol{x} \in X_t} \boldsymbol{x} \\ &= \sum_{t \in T} \frac{|X_t|}{|X|} \cdot \frac{1}{|X_t|} \sum_{\boldsymbol{x} \in X_t} \boldsymbol{x} \\ &= \sum_{t \in T} p_t \mathbb{E}_{\boldsymbol{x} \in X_t} \{\boldsymbol{x}\} \\ &= \sum_{t \in T} p_t \boldsymbol{\mu}(X_t). \end{aligned}$$

□

E.3 Calculation of $Q(X)$

By replacing X_t with X in $Q(X_t)$ in (4), $Q(X)$ for X is defined. We prove (15) given by:

$$Q(X) = \mathbb{E}_{\boldsymbol{x} \in X}\left\{\|\boldsymbol{x}\|^2\right\} = \sum_{t \in T} p_t Q(X_t).$$

To do so, we first prove the following equation for $q_i(X)$, where $q_i(X)$ is the value obtained by replacing X_t with X in $q_i(X_t)$ as defined in (17) in Appendix C:

$$q_i(X) = \sum_{t \in T} p_t q_i(X_t). \tag{25}$$

Proof.

$$\begin{aligned}
q_i(X) &= \mathbb{E}_{\boldsymbol{x} \in X}\{x_i^2\} \\
&= \frac{1}{|X|} \sum_{\boldsymbol{x} \in X} x_i^2 \\
&= \frac{1}{|X|} \sum_{t \in T} \sum_{\boldsymbol{x} \in X_t} x_i^2 \\
&= \sum_{t \in T} \frac{|X_t|}{|X|} \cdot \frac{1}{|X_t|} \sum_{\boldsymbol{x} \in X_t} x_i^2 \\
&= \sum_{t \in T} p_t \mathbb{E}_{\boldsymbol{x} \in X_t}\{x_i^2\} \\
&= \sum_{t \in T} p_t q_i(X_t).
\end{aligned}$$

□

Using (25), we then show (15):

$$\begin{aligned}
Q(X) &= \sum_{i=1}^{d} q_i(X) \quad \text{(from (21), where } X_t \text{ is replaced by } X) \\
&= \sum_{i=1}^{d} \sum_{t \in T} p_t q_i(X_t) \quad (\because (25)) \\
&= \sum_{t \in T} p_t \sum_{i=1}^{d} q_i(X_t) \\
&= \sum_{t \in T} p_t Q(X_t) \quad (\because (21)).
\end{aligned}$$

F Relationships between frequency and $Q(X_t)$, $M(X_t)$, and $V(X_t)$

Similar to previous work (Wannasuphoprasit et al., 2023; Liang et al., 2021; Zhou et al., 2021, 2022a), We shows scatter plots[9] that represent the relationships between frequency and the values of $Q(X_t)$, $M(X_t)$, and $V(X_t)$ for some layers of BERT in Fig. 11, RoBERTa in Fig. 12, and GPT-2 in Fig. 13, respectively. Regression lines are also shown for these plots. The slopes of the lines for $Q(X_t)$ increase as the layers deepen, while those for $M(X_t)$ and $V(X_t)$ are consistently negative and positive, respectively. These trends are consistent with those observed in previous work. Additionally, the rightmost columns of

[9]Unlike previous work, $\log_{10}$ scales are used for $Q(X_t)$, $M(X_t)$, and $V(X_t)$ to address the large value differences.

Figs. 11, 12, and 13 show the scatter plots and regression lines for $M(X_t)$ and $V(X_t)$. In the intermediate layer, the slope is smaller and the R^2 value is larger than that of other layers. These results indicate a strong trade-off relationship between $M(X_t)$ and $Q(X_t)$ in the intermediate layer.

Additionally, histograms of $Q(X_t)$, $M(X_t)$, and $V(X_t)$ for the layers of the models are shown in Fig. 14 for BERT, Fig. 15 for RoBERTa, and Fig. 16 for GPT-2.

Based on Figs. 11, 12, and 13, Fig. 17 shows the slopes of the regression lines for frequency and $Q(X_t)$, $M(X_t)$, and $V(X_t)$ across the six models and layers. The slopes for $Q(X_t)$ remain stable and close to 0 across layers, although an increase is observed in GPT-2. As the layers deepen, the slopes for $M(X_t)$ and $V(X_t)$ remain approximately negative and positive across all layers. The variation in slopes remains small and stable for all values.

G Detailed results of the statistical measures for X

In this section, we present detailed results from Section 5.3. Using the data from the bar graphs in Fig. 5, Fig. 18 shows the normalized values of $M(X)$, $V(X)(= V_W(X) + V_B(X))$, $V_W(X)$, and $V_B(X)$ relative to $Q(X)$ for each layer of each model. As observed in Fig. 5, $M(X)/Q(X)$ increases as the layers deepen, while $V(X)/Q(X)$ decreases. Furthermore, $V_W(X)/Q(X)$ increases, whereas $V_B(X)/Q(X)$ decreases. Figure 19 shows the values of $Q(X)$, $M(X)$, and $V(X)$ for each layer across the six models. As seen in Fig. 18, where $M(X)/Q(X)$ increases as the layers deepen, we observe that $M(X)$ also increases, although the increase varies among the models. In contrast, while $V(X)/Q(X)$ decreases in GPT-2 as the layers deepen, the value of $V(X)$ itself increases monotonically, except in the final layer. It is known that in GPT-2, the norm and standard deviation of the residual stream increase exponentially as the layers deepen (Heimersheim and Turner, 2023), and the results in Fig. 19 are consistent with previous work.

Figure 20 shows the values of $V_W(X)$, $V_B(X)$, and $V_B(X)/V(X)$ for each layer across the six models. As observed in Fig. 6, where $V_W(X)/V(X)$ increases as the layers deepen, we can also see that the value of $V_W(X)$ increases. In contrast, in GPT-2, $V_B(X)/V(X)$ decreases as the layers deepen, while the value of $V_B(X)$ itself increases monotonically, except in the final layer.

H Explanation of the differences in Transformer layers

This section introduces the Transformer layers based on the explanation by Kobayashi et al. (2024). A single Transformer layer (Vaswani et al., 2017) consists of four components: multi-head attention (ATTN), feed-forward network (FF), residual connection (RES), and layer normalization (LN), as shown in Fig. 21. Following Xiong et al. (2020); Kobayashi et al. (2024), we classify the layers into Post-LN and Pre-LN Transformer layers based on the position of the LN. In the models we used, BERT and RoBERTa have Post-LN layers, while GPT-2 has Pre-LN layers.

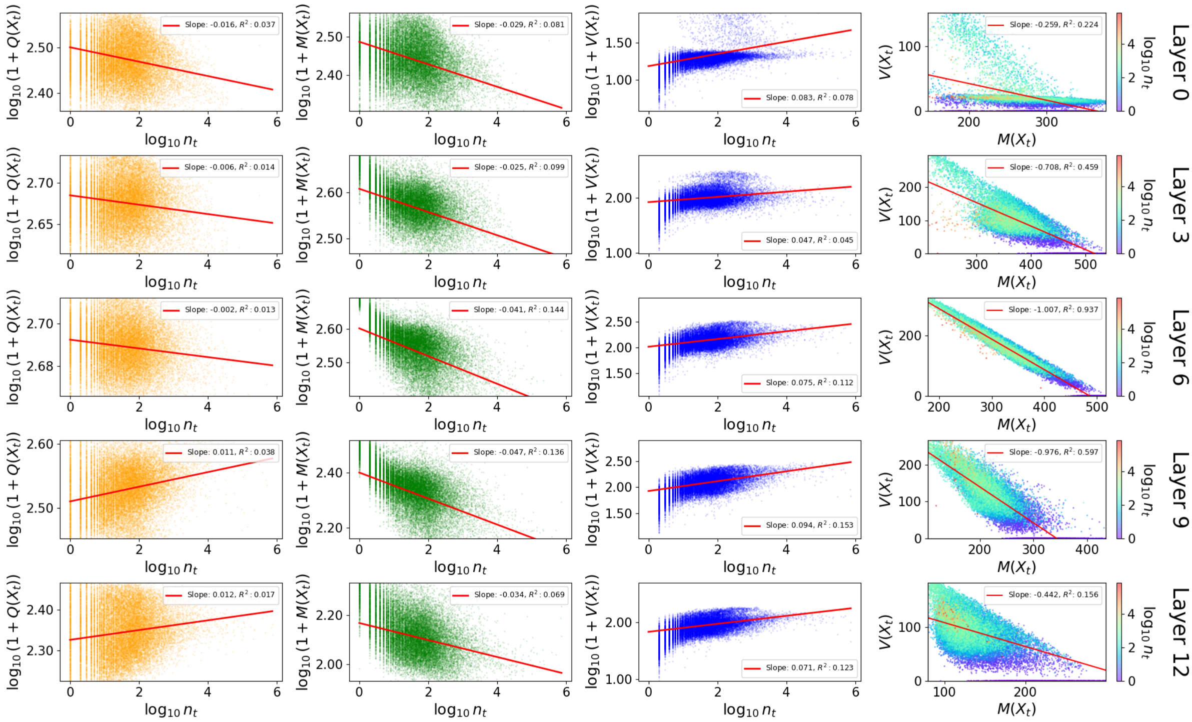
A single Transformer layer can be divided into two parts: the Attention Block (ATB), consisting of ATTN, RES1, and LN1, and the Feed-Forward Block (FFB), consisting of FF, RES2, and LN2. In this study, we focus on the FFB because we are analyzing the output of each layer. For a detailed explanation of the ATB, see Kobayashi et al. (2024). The FF, RES, and LN functions take $\boldsymbol{h} \in \mathbb{R}^d$ as input and are defined as follows:

$$FF(\mathbf{h}) = \boldsymbol{W}_2\, \boldsymbol{g}(\boldsymbol{W}_1 \boldsymbol{h} + \boldsymbol{b}_1) + \boldsymbol{b}_2 \in \mathbb{R}^d, \tag{26}$$

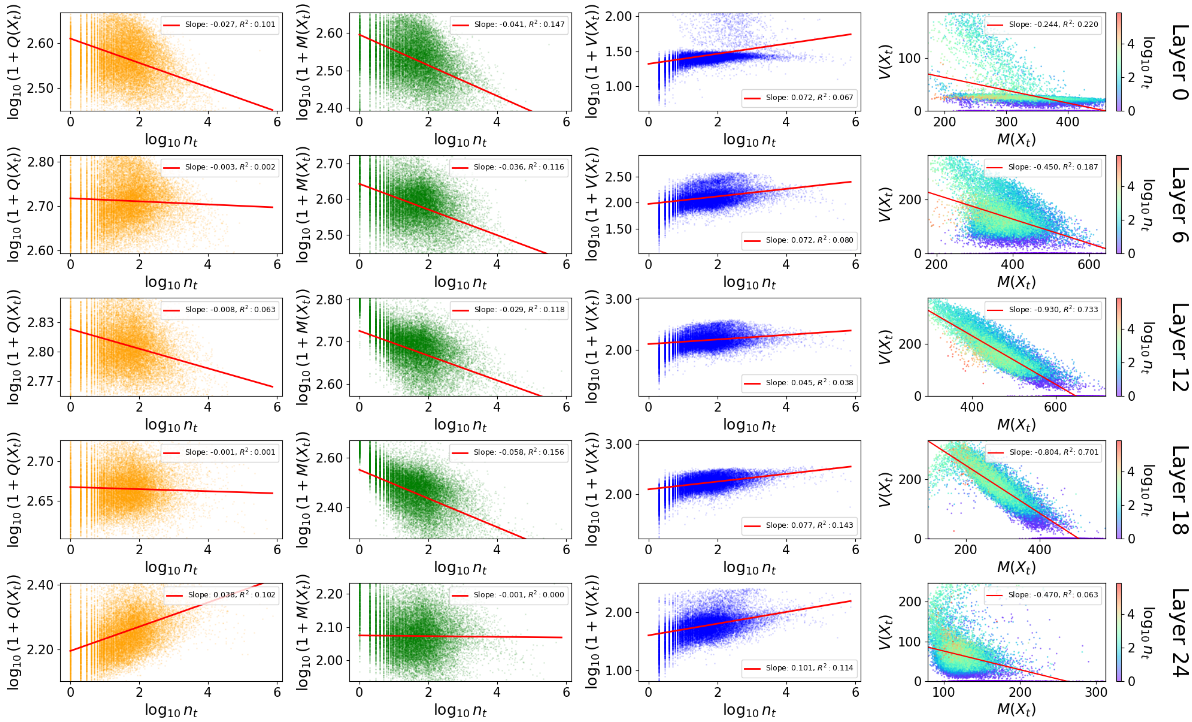
$$(RES \circ \boldsymbol{v})(\mathbf{h}) = \boldsymbol{v}(\mathbf{h}) + \mathbf{h} \in \mathbb{R}^d, \tag{27}$$

$$LN(\mathbf{h}) = \frac{\boldsymbol{h} - \text{Mean}(\boldsymbol{h})}{\text{Std}(\boldsymbol{h})} \odot \boldsymbol{\gamma} + \boldsymbol{\beta} \in \mathbb{R}^d, \tag{28}$$

where $\boldsymbol{W}_1 \in \mathbb{R}^{d' \times d}$ and $\boldsymbol{b}_1 \in \mathbb{R}^{d'}$ are the weight and bias of the input layer in the FF, $\boldsymbol{W}_2 \in \mathbb{R}^{d \times d'}$ and $\boldsymbol{b}_2 \in \mathbb{R}^d$ are the weight and bias of the output layer in the FF, and $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are the weight and bias of the LN. In addition, $\boldsymbol{g} : \mathbb{R}^{d'} \mapsto \mathbb{R}^{d'}$, $\boldsymbol{v} : \mathbb{R}^d \mapsto \mathbb{R}^d$, $\text{Mean} : \mathbb{R}^d \mapsto \mathbb{R}$, and $\text{Std} : \mathbb{R}^d \mapsto \mathbb{R}$ represent the activation function, arbitrary vector-valued functions, the mean of the elements, and the standard deviation of the elements, respectively. The operator $\odot$ denotes element-wise multiplication.

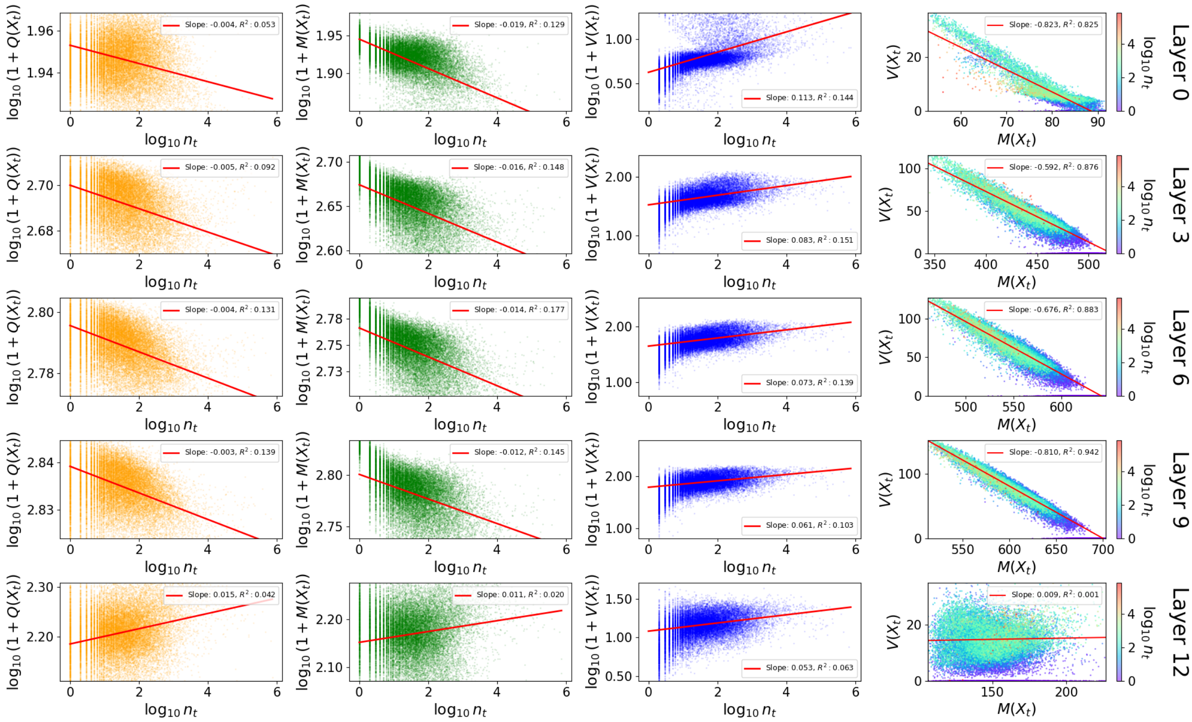


(a) bert-base-uncased

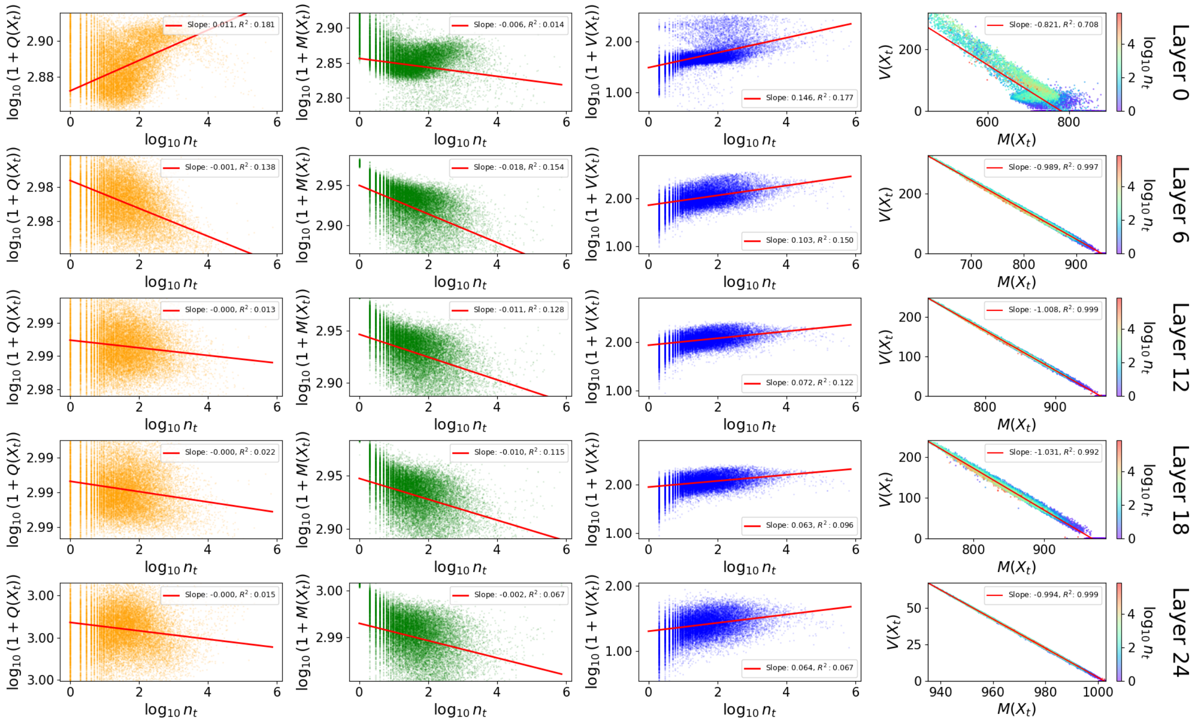


(b) bert-large-uncased

Figure 11: Scatter plots across some layers from the input to the output layer of (a) `bert-base-uncased` and (b) `bert-large-uncased`, plotting $\log_{10}(1 + Q(X_t))$, $\log_{10}(1 + M(X_t))$, and $\log_{10}(1 + V(X_t))$ against $\log_{10} n_t$, and plotting $V(X_t)$ against $M(X_t)$. Each plot includes a regression line, its slope, and the coefficient of determination, R^2. Only tokens with $1 \leq \log_{10} n_t \leq 5$ were used for regressions to reduce the influence of extreme values.

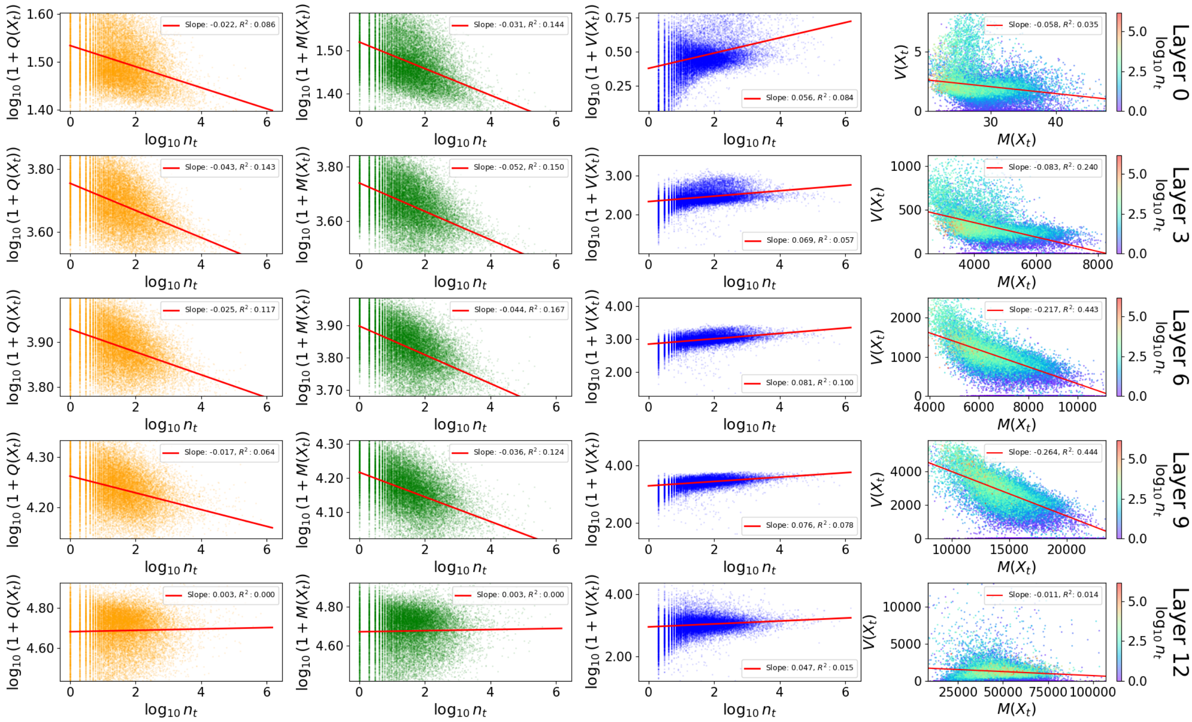


(a) roberta-base

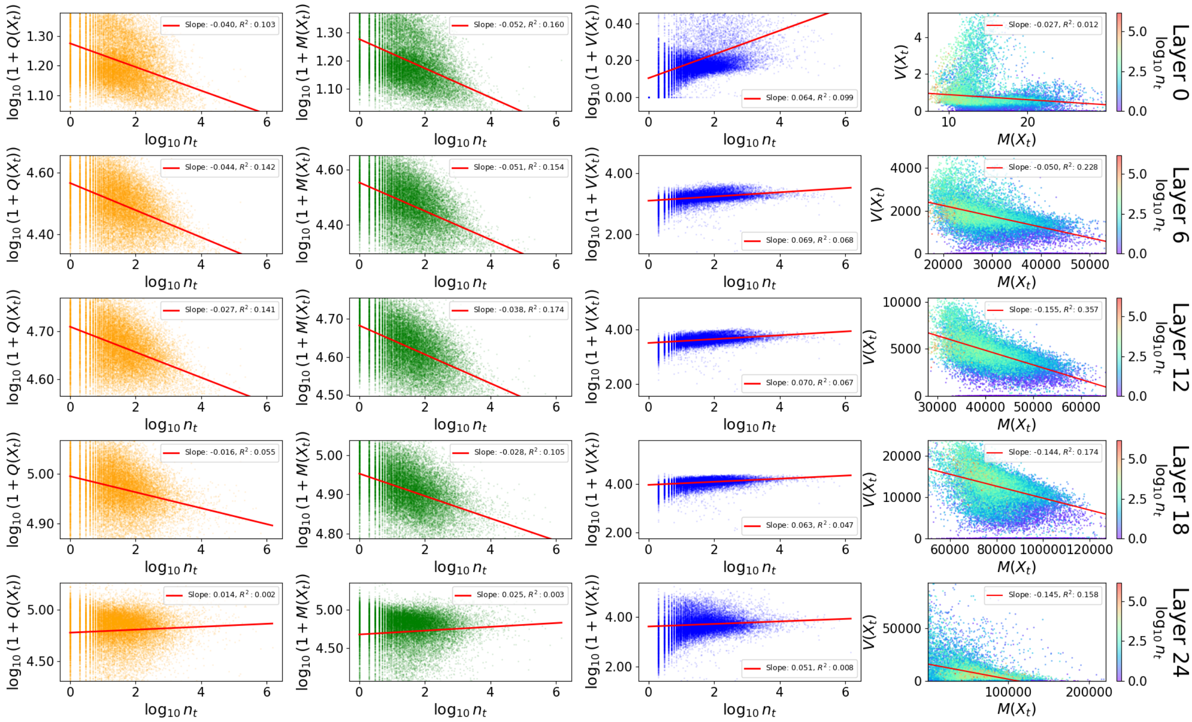


(b) roberta-large

Figure 12: Results of the same experiments as in Fig. 11 for (a) `roberta-base` and (b) `roberta-large`.

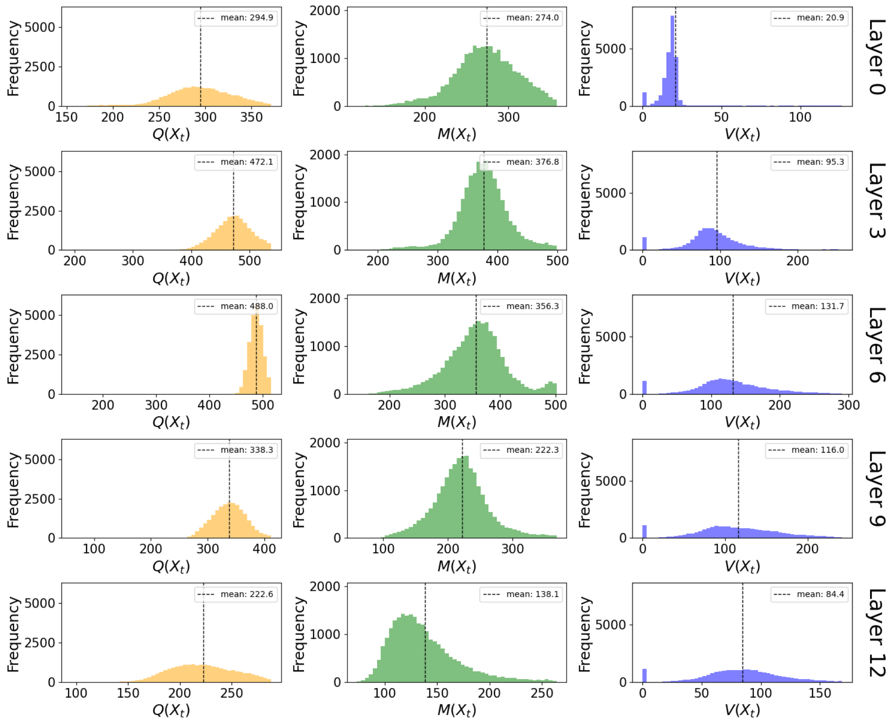


(a) gpt2

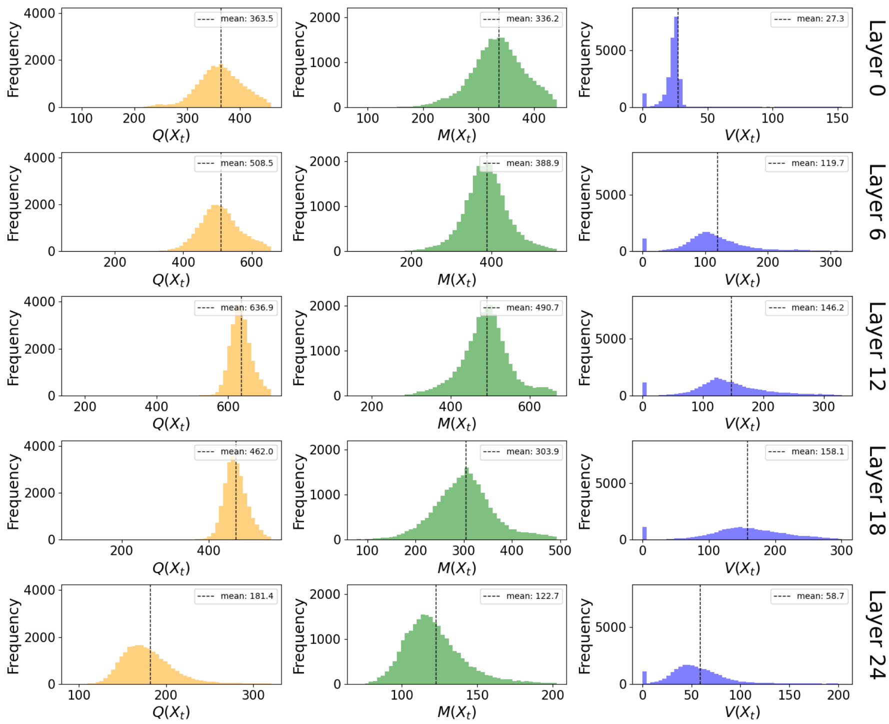


(b) gpt2-medium

Figure 13: Results of the same experiments as in Fig. 11 for (a) `gpt2` and (b) `gpt2-medium`.

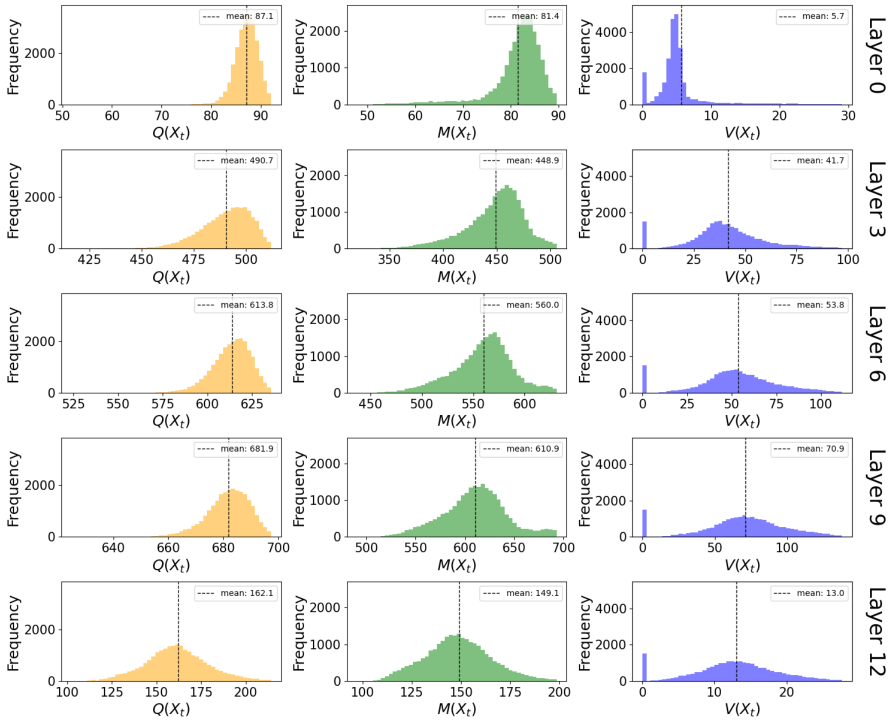


(a) bert-base-uncased

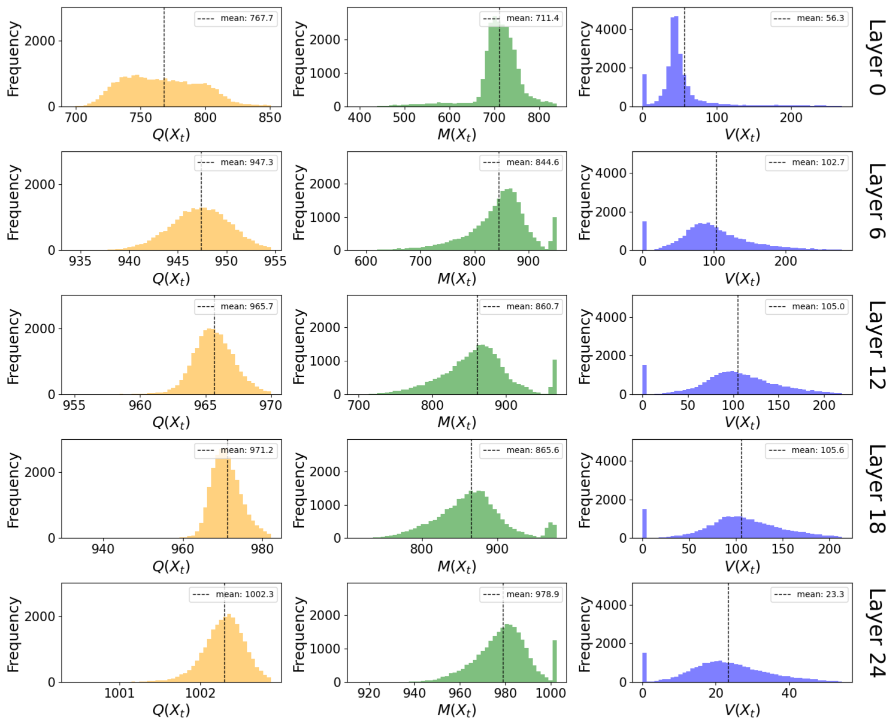


(b) bert-large-uncased

Figure 14: Histograms of $Q(X_t)$, $M(X_t)$, and $V(X_t)$ for each layer of (a) `bert-base-uncased` and (b) `bert-large-uncased`.

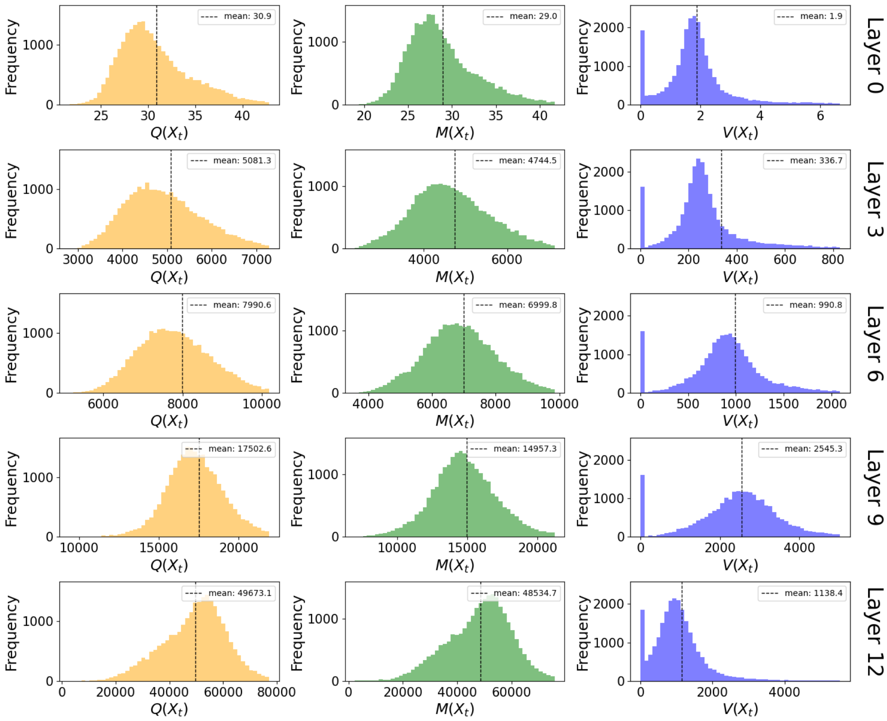


(a) roberta-base

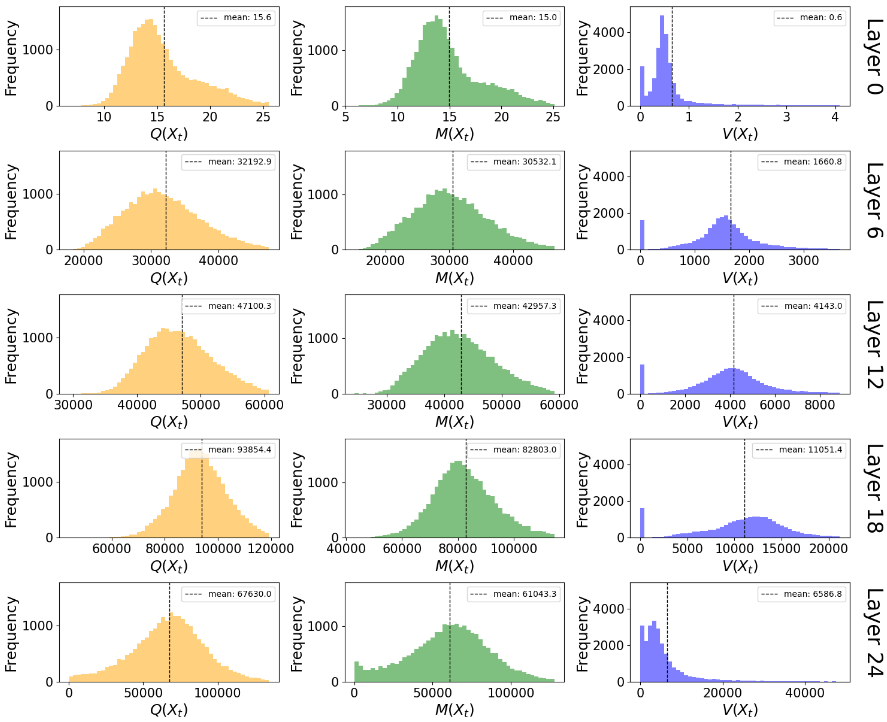


(b) roberta-large

Figure 15: Histograms of $Q(X_t)$, $M(X_t)$, and $V(X_t)$ for each layer of (a) `roberta-base` and (b) `roberta-large`.



(a) gpt2



(b) gpt2-medium

Figure 16: Histograms of $Q(X_t)$, $M(X_t)$, and $V(X_t)$ for each layer of (a) `gpt2` and (b) `gpt2-medium`.

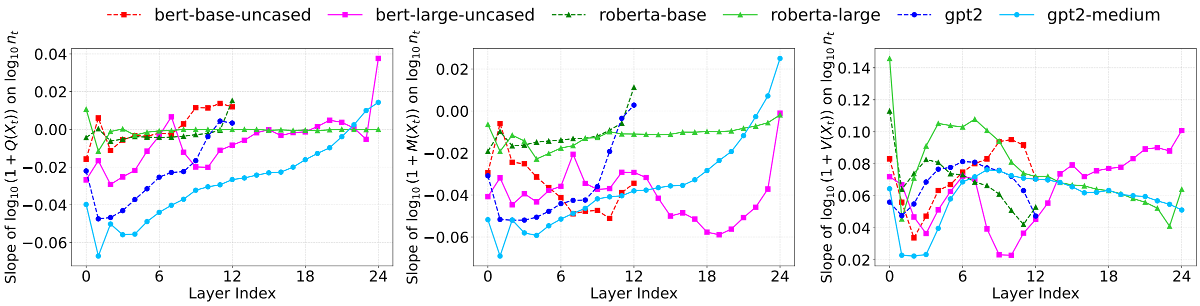


Figure 17: Regression slopes between $\log_{10}(n_t)$ and each of $\log_{10}(1+Q(X_t))$, $\log_{10}(1+M(X_t))$, and $\log_{10}(1+V(X_t))$ across the six models and layers. As the layers deepen, the slopes for $Q(X_t)$ tend to increase. Across all layers, the slopes for $M(X_t)$ generally remain negative, while those for $V(X_t)$ remain positive. Only tokens with $1 \leq \log_{10} n_t \leq 5$ were used for regressions to reduce the influence of extreme values.

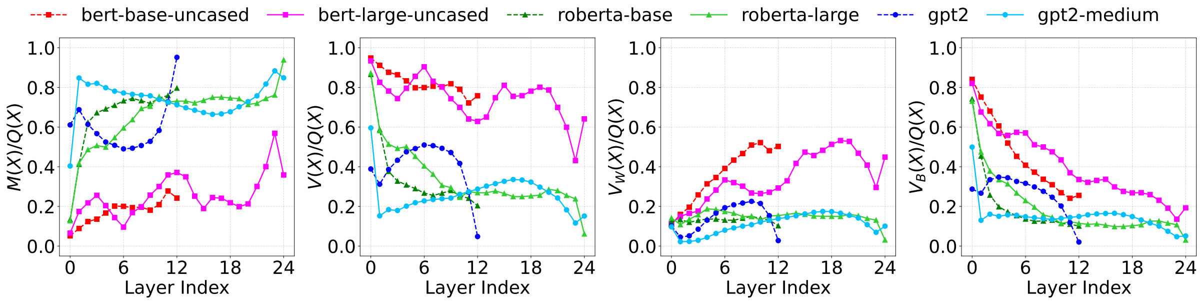


Figure 18: The values of $M(X)$, $V(X)(= V_W(X)+V_B(X))$, $V_W(X)$, and $V_B(X)$ normalized by $Q(X)$ for each layer of each model, based on Fig. 5.

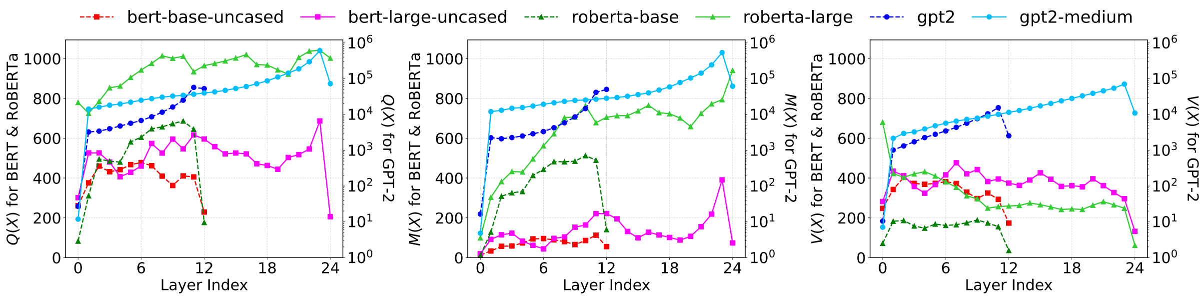


Figure 19: Values of $Q(X)$, $M(X)$, and $V(X)$ for each layer across the six models. For GPT-2, refer to the right vertical axis, as the scale of the values differs from those of BERT and RoBERTa.

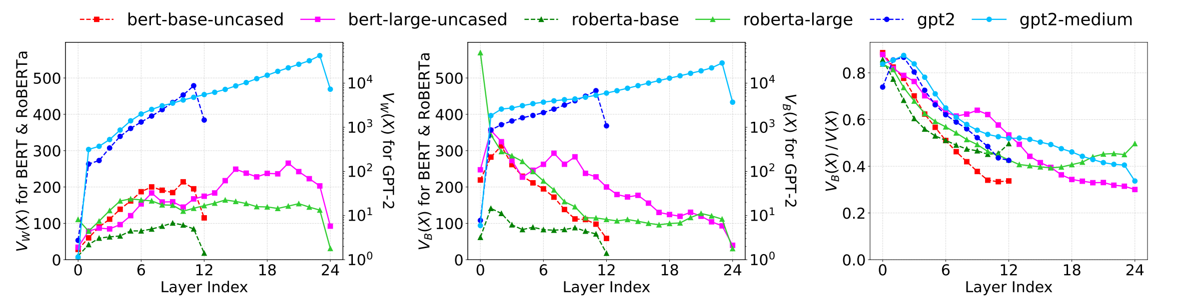


Figure 20: Values of $V_W(X)$, $V_B(X)$, and $V_B(X)/V(X)$ for each layer across the six models. For GPT-2, refer to the right vertical axis for $V_W(X)$ and $V_B(X)$, as the scale of the values differs from those of BERT and RoBERTa. Since $V_B(X)/V(X) = 1 - V_W(X)/V(X)$, similar to Fig. 6, where $V_W(X)/V(X)$ increases as the layers deepen, $V_B(X)/V(X)$ decreases.

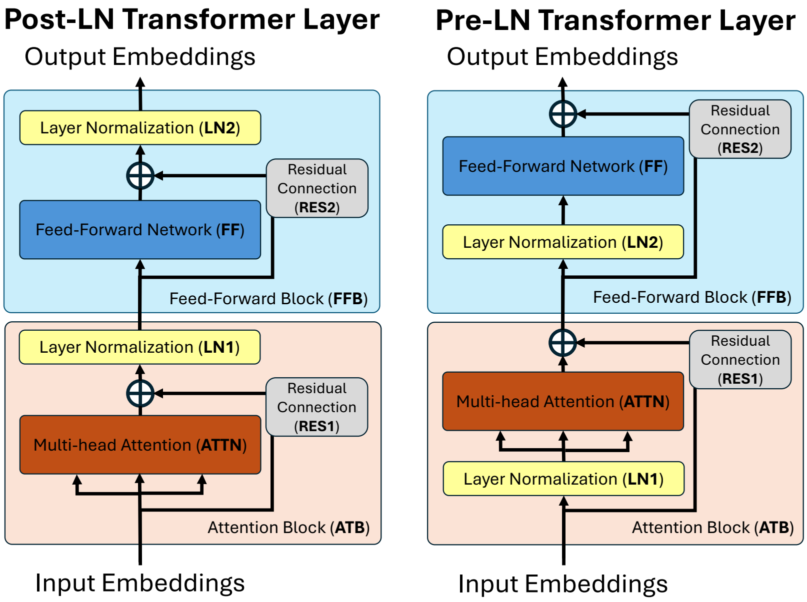


Figure 21: Figure of the Post-LN and Pre-LN Transformer layer based on Kobayashi et al. (2024). BERT and RoBERTa have Post-LN layers, while GPT-2 has Pre-LN layers.

We denote the FFB of the Post-LN Transformer layer as $\mathrm{FFB}_{\mathrm{Post}}$, that of the Pre-LN Transformer layer as $\mathrm{FFB}_{\mathrm{Pre}}$, and the output of the ATB as $\boldsymbol{h}_{\mathrm{ATB}} \in \mathbb{R}^d$, then we have:

$$\mathrm{FFB}_{\mathrm{Post}}(\boldsymbol{h}_{\mathrm{ATB}}) = (\mathrm{LN2} \circ \mathrm{RES2} \circ \mathrm{FF})(\boldsymbol{h}_{\mathrm{ATB}}), \tag{29}$$

$$\mathrm{FFB}_{\mathrm{Pre}}(\boldsymbol{h}_{\mathrm{ATB}}) = (\mathrm{RES2} \circ \mathrm{FF} \circ \mathrm{LN2})(\boldsymbol{h}_{\mathrm{ATB}}). \tag{30}$$

I Variation of $Q(X_t)$ for the embeddings from the Layer Normalization

Following the definition in (28), we consider the case that embedding is expressed as

$$\begin{aligned}\boldsymbol{x} &= LN(\boldsymbol{h}) \\ &= \boldsymbol{z} \odot \boldsymbol{\gamma} + \boldsymbol{\beta} \in \mathbb{R}^d,\end{aligned}$$

where

$$\boldsymbol{z} := \frac{\boldsymbol{h} - \mathrm{Mean}(\boldsymbol{h})}{\mathrm{Std}(\boldsymbol{h})}. \tag{31}$$

Corresponding to the sampling $\boldsymbol{x} \in X_t$ for the token-wise embedding set, we define the sampling $\boldsymbol{z} \in Z_t$, and assume that Z_t is sampled from a distribution F_t. Thus

$$\mathbb{E}_{\boldsymbol{z} \in Z_t}\{z_i^k\} = \mathbb{E}_{\boldsymbol{z} \sim F_t}\{z_i^k\} + O_p(n_t^{-1/2}), \tag{32}$$

where $n_t = |Z_t|$ is the sample size. Here we introduce an assumption that the marginal distributions of the elements $z_1, \ldots, z_d$ of $\boldsymbol{z} \sim F_t$ are the same; Although this setting does not necessarily reflect reality, we assume it as an ideal scenario. Since (31), the sample mean and the sample variance of the elements $z_1, \ldots, z_d$ are zero and one, respectively, we can assume that, for a sufficiently large d, the population mean and the population variance of each element z_i in F_t is also zero and one, respectively. Therefore, (32) with $k = 0$ and $k = 1$ gives

$$\mathbb{E}_{\boldsymbol{z} \in Z_t}\{z_i\} = O_p(n_t^{-1/2}), \quad \mathbb{E}_{\boldsymbol{z} \in Z_t}\{z_i^2\} = 1 + O_p(n_t^{-1/2}).$$

Now we look at $Q(X_t)$ for the embeddings from the layer normalization.

$$\begin{aligned}
Q(X_t) &= \mathbb{E}_{\boldsymbol{x}\in X_t}\{\|\boldsymbol{x}\|^2\} \\
&= \mathbb{E}_{\boldsymbol{z}\in Z_t}\{\|\boldsymbol{z}\odot\boldsymbol{\gamma}+\boldsymbol{\beta}\|^2\} \\
&= \mathbb{E}_{\boldsymbol{z}\in Z_t}\left\{\sum_{i=1}^{d}(\gamma_i z_i+\beta_i)^2\right\} \\
&= \sum_{i=1}^{d}\left\{\gamma_i^2\mathbb{E}_{\boldsymbol{z}\in Z_t}\{z_i^2\}+2\beta_i\gamma_i\mathbb{E}_{\boldsymbol{z}\in Z_t}\{z_i\}+\beta_i^2\right\} \\
&= \sum_{i=1}^{d}\left\{\gamma_i^2(1+O_p(n_t^{-1/2}))+2\beta_i\gamma_i O_p(n_t^{-1/2})+\beta_i^2\right\} \\
&= \|\boldsymbol{\gamma}\|^2+\|\boldsymbol{\beta}\|^2+(\|\boldsymbol{\gamma}\|^2+\boldsymbol{\beta}^\top\boldsymbol{\gamma})O_p(n_t^{-1/2}).
\end{aligned}$$

This implies that $Q(X_t)\approx\|\boldsymbol{\gamma}\|^2+\|\boldsymbol{\beta}\|^2$ is nearly constant, with variation proportional to $n_t^{-1/2}$. Since we evaluate the variation of $Q(X_t)$ when sampling $t\in T$, the worst case is $n_0=\min_{t\in T} n_t$. Therefore, the coefficient of variation (C.V.) is

$$\text{C.V.}(Q(X_t))=\frac{(\|\boldsymbol{\gamma}\|^2+\boldsymbol{\beta}^\top\boldsymbol{\gamma})O(n_0^{-1/2})}{\|\boldsymbol{\gamma}\|^2+\|\boldsymbol{\beta}\|^2}=O(n_0^{-1/2}).$$

This C.V. approaches zero as both d and all n_t become larger.

J Artifacts of word embeddings represented by the mean of token embeddings

In this study, we used token embeddings in our experiments. In contrast, some previous work investigating the relationship between frequency and embeddings has used **word** embeddings represented by the mean of token embeddings (e.g., Wannasuphoprasit et al. (2023)). This section explains the artifacts that can arise when using such word embeddings.

Let S_w be the set of sentences containing the word w in the corpus, and T_w be the set of tokens when word w is tokenized. Similar to the token embedding set X_t defined in (2), we define the embedding set of word w using the d-dimensional contextualized embedding model f and S_w as follows:

$$X_w := \left\{\frac{1}{|T_w|}\sum_{t\in T_w} f(s,t)\;\middle|\; s\in S_w\right\}. \tag{33}$$

Using `bert-base-uncased` as the embedding model, we performed the same experiments on X_w as those in Fig. 11a, and we show the results in Fig. 22, with words colored by $|T_w|$. Here, unlike Fig. 11a, we plotted the actual values of $Q(X_t)$, $M(X_t)$, and $V(X_t)$ instead of using a log scale for better visualization. As shown in Fig. 22, words with the same number of tokens tend to form clusters, especially in the shallow layers. Additionally, the values of $Q(X_w)$, $M(X_w)$, and $V(X_w)$ become smaller as the number of tokens increases. This is likely because, as the number of tokens increases, the averaged component values tend to approach zero. These results are consistent with previous research (Zhou et al., 2022b), which showed that the mean norm tends to become smaller as the number of subwords increases. With this in mind, we conducted our analysis in this study using token embeddings.

Interestingly, this artifact diminishes as the embeddings become more contextualized in the deeper layers. Therefore, in the experiments conducted by Wannasuphoprasit et al. (2023), where only the final layer of `bert-base-uncased` was used to analyze $\mathbb{E}_{\boldsymbol{x}\in X_w}\{\|\boldsymbol{x}\|\}$, the effect of such artifacts appeared to be relatively small.

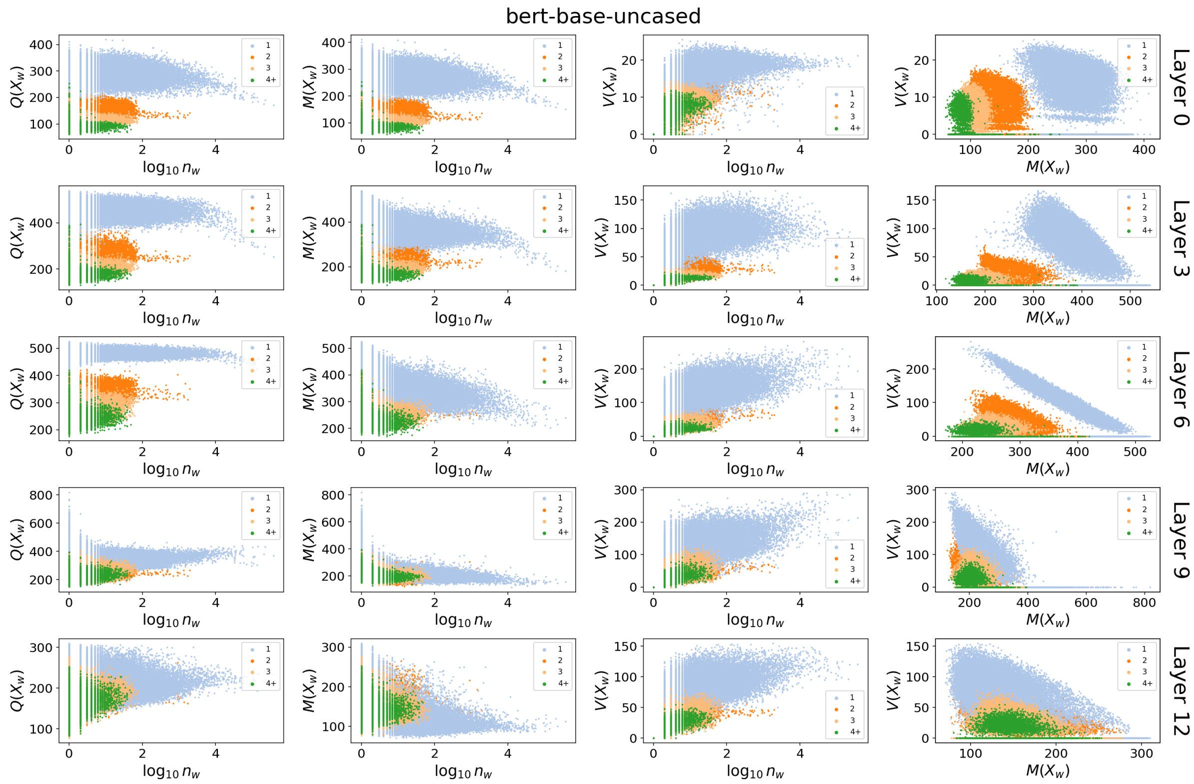


Figure 22: Results of the same experiments as in Fig. 11a using each word embedding set X_w, where the embeddings are the means of the token embeddings for `bert-base-uncased`. Words are colored based on the number of tokens produced by BERT tokenization. The clustering of words with the same color indicates an artifact caused by taking the means of the token embeddings.