frontiers | Frontiers in Artificial Intelligence

TYPE Methods
PUBLISHED 17 September 2025
DOI 10.3389/frai.2025.1663484

OPEN ACCESS

EDITED BY
Mini Han Wang,
Zhuhai People's Hospital, China

REVIEWED BY
Antonio Sarasa-Cabezuelo,
Complutense University of Madrid, Spain
Hunter Scarborough,
John Peter Smith Hospital, United States
Hojjat Karami,
Swiss Federal Institute of Technology Lausanne, Switzerland

*CORRESPONDENCE
Ali Amirahmadi
ali.amirahmadi@hh.se

RECEIVED 10 July 2025
ACCEPTED 25 August 2025
PUBLISHED 17 September 2025

CITATION
Amirahmadi A, Etminani F and Ohlsson M (2025) Adaptive noise-augmented attention for enhancing Transformer fine-tuning on longitudinal medical data.
Front. Artif. Intell. 8:1663484.
doi: 10.3389/frai.2025.1663484

Adaptive noise-augmented attention for enhancing Transformer fine-tuning on longitudinal medical data

Ali Amirahmadi[1]*, Farzaneh Etminani[1,2] and Mattias Ohlsson[3]

[1]Center for Applied Intelligent Systems Research in Health, Halmstad University, Halmstad, Sweden, [2]Department of Research and Development (FoU), Region Halland, Halmstad, Sweden, [3]Centre for Environmental and Climate Science, Computational Science for Health and Environment, Lund University, Lund, Sweden

Transformer models pre-trained on self-supervised tasks and fine-tuned on downstream objectives have achieved remarkable results across a variety of domains. However, fine-tuning these models for clinical predictions from longitudinal medical data, such as electronic health records (EHR), remains challenging due to limited labeled data and the complex, event-driven nature of medical sequences. While self-attention mechanisms are powerful for capturing relationships within sequences, they may underperform when modeling subtle dependencies between sparse clinical events under limited supervision. We introduce a simple yet effective fine-tuning technique, Adaptive Noise-Augmented Attention (ANAA), which injects adaptive noise directly into the self-attention weights and applies a 2D Gaussian kernel to smooth the resulting attention maps. This mechanism broadens the attention distribution across tokens while refining it to emphasize more informative events. Unlike prior approaches that require expensive modifications to the architecture and pre-training phase, ANAA operates entirely during fine-tuning. Empirical results across multiple clinical prediction tasks demonstrate consistent performance improvements. Furthermore, we analyze how ANAA shapes the learned attention behavior, offering interpretable insights into the model's handling of temporal dependencies in EHR data.

KEYWORDS

Transformer, augmentation, adaptive noise, medical data, electronic health records (EHR), fine-tuning, representation learning, self-attention

1 Introduction

Foundation models, deep neural networks pre-trained on broad unlabeled data using self-supervised methods, have significantly impacted various aspects of our lives, including law, healthcare, education, and more (Bommasani et al., 2021; Guo et al., 2023; Wornow et al., 2023). These models typically acquire general knowledge about the data through pre-training a variant of the Transformer network on a self-supervised task like Masked Language Model (MLM), and then adapt this knowledge to downstream tasks with only a few labeled samples during the fine-tuning process. Researchers showed that pre-training, even with limited data, can improve Transformers' performance significantly (Amos et al., 2023).

Pre-training Transformers have been employed with various self-supervised objectives and domains. Common objectives include corrupted text reconstruction tasks like MLM (Devlin et al., 2018; Lewis et al., 2019; Lan et al., 2019) and standard language models

such as next-word prediction (Radford et al., 2019; Brown et al., 2020), which have been extensively utilized (Liu et al., 2023). These models typically adopt a backbone architecture inspired by the multi-head attention mechanism in Transformers (Vaswani et al., 2017), known for its effectiveness in modeling complex interaction between events (tokens) in a sequence (text). These foundation models have been pre-trained on different domain data (Lan et al., 2019; Radford et al., 2019), including structured temporal health data as sequences of events (Li et al., 2020; Rasmy et al., 2021; Pang et al., 2021).

Modeling Electronic Health Records (EHRs) trajectories presents a critical opportunity for predicting health-related outcomes, offering benefits like early intervention, cost reduction, and improved public health. This field has attracted significant attention from deep learning researchers (Xiao et al., 2018; Amirahmadi et al., 2023; Boll et al., 2024; Li et al., 2024). Typically, healthcare specific foundation models are pre-trained on publicly available, unlabeled EHR data, and adapting these models through fine-tuning consistently demonstrates superior performance across various tasks (Li et al., 2020; Rasmy et al., 2021; Pang et al., 2021; Ren et al., 2021; Li et al., 2022; Yuanyuan et al., 2025).

However, EHRs are often scarce, and training Transformers to learn the complex relationships between medical events in longitudinal EHRs requires either large amounts of data, or advanced training techniques and augmentations (Dosovitskiy et al., 2020; Touvron et al., 2021; Hassani et al., 2021, 2023). Due to privacy concerns and the scarcity of publicly available datasets, models often fail to learn the intricate dependencies between events in a patient's history. To address this, Choi et al. (2020) proposed incorporating domain knowledge into the attention mechanism, while Zhu and Razavian (2021) employed variational regularization. Additionally, Amirahmadi et al. (2025) suggested pre-training the Transformer on the MLM task and the ordering of medical events in a patient's history, and Kim and Lee (2024) proposed using learnable, adaptive kernels in the attention matrices to improve contextual representations and enhance the learned structure through self-attention. Figures 1, 4 illustrate how these various approaches impact self-attention behaviors in leaning the relationships between events. However, these methods often come with substantial computational costs and require extra effort for implementation and design.

Data augmentation is another solution to tackle the data scarcity challenge. Augmenting data with discrete data types, such as series of medical codes or tokens in text, is challenging because small perturbations can drastically alter semantic meaning, and interpolation in discrete space is not feasible (Chen et al., 2020). For example, replacing a code for "Type 1 diabetes" with "Type 2 diabetes," or reordering diagnosis and procedure codes within the same patient trajectory, can fundamentally change the clinical context. As a result, researchers have proposed augmenting models during training as an alternative (Jain et al., 2023; Zehui et al., 2019; Wu et al., 2023).

In this study, we propose a simple two-step augmentation technique-Adaptive Noise-Augmented Attention (ANAA)—that perturbs attention scores by injecting adaptive Gaussian noise followed by smoothing with a Gaussian kernel. Our investigation of attention distributions reveals that fine-tuned Transformers tend to produce highly polarized attention scores—values clustering near the extremes (0 or 1), which restricts the model's capacity to explore diverse dependencies (see the bottom row of Figure 4). By introducing controlled noise into attention scores during fine-tuning, we encourage exploration of alternative dependency paths between events. The subsequent smoothing operation helps restore structural consistency while preserving diversity, resulting in more balanced and informative self-attention maps.

The main contributions are summarized as follows:

1. We proposed a simple self-attention augmentation method that encourages the model to explore and learn more complex attention patterns during fine-tuning. Importantly, this approach does not modify the computational graph, making it easily applicable to any pre-trained Transformer.
2. We conducted several evaluations on various downstream tasks, examining the effect of the novel method on model performance, model robustness with limited training samples, and the balance of attention distribution between distant and nearby events. Our results demonstrate how it improves the performance of pre-trained Transformers.

2 Preliminary

2.1 Transformer encoder and self-attention

The core back-bone of Transformers encoder is the multi-head self-attention. Each self-attention head is:

$$Q_h = XW_h^Q, K_h = XW_h^K, V_h = XW_h^V, \tag{1}$$

$$A_h = \text{softmax}(\frac{Q_h K_h^T}{\sqrt{d_k}}) \tag{2}$$

$$H_h = \text{Self-attention}(X) = A_h V_h \tag{3}$$

Where, $Q, K \in \mathbb{R}^{n \times d_k}$ and $V \in \mathbb{R}^{n \times d_v}$ and n is the length of input sequence and d_k and d_v are dimenssion of Key and Value. A_h is the attention score matrix and each $A_{i,j}$ indicates how much attention token x_i put on x_j. Transformer encoders, is built on concatenation of $|\ h\ |$ number attention heads in parallel, so each one has its own weights. Then, the concatenation is projected:

$$\text{MultiHead}(X) = \text{Concat}(H_1 ...., H_{|h|}) W^O \tag{4}$$

Where, $W^O \in \mathbb{R}^{|h| \times d_v}$ Multiple self-attention heads in parallel, help the model to attend to information from different representation subspaces (Vaswani et al., 2017; Hao et al., 2021).

2.2 Pre-training, fine-tuning

Pretraining typically involves the model acquiring general knowledge, which is then used to initialize the final network. Subsequently, the final network adjusts these weights to obtain optimized weights for specific downstream tasks (Chen et al., 2021). This approach has been extensively utilized for adapting foundation models to downstream tasks (Lan et al., 2019; Liu et al., 2023).

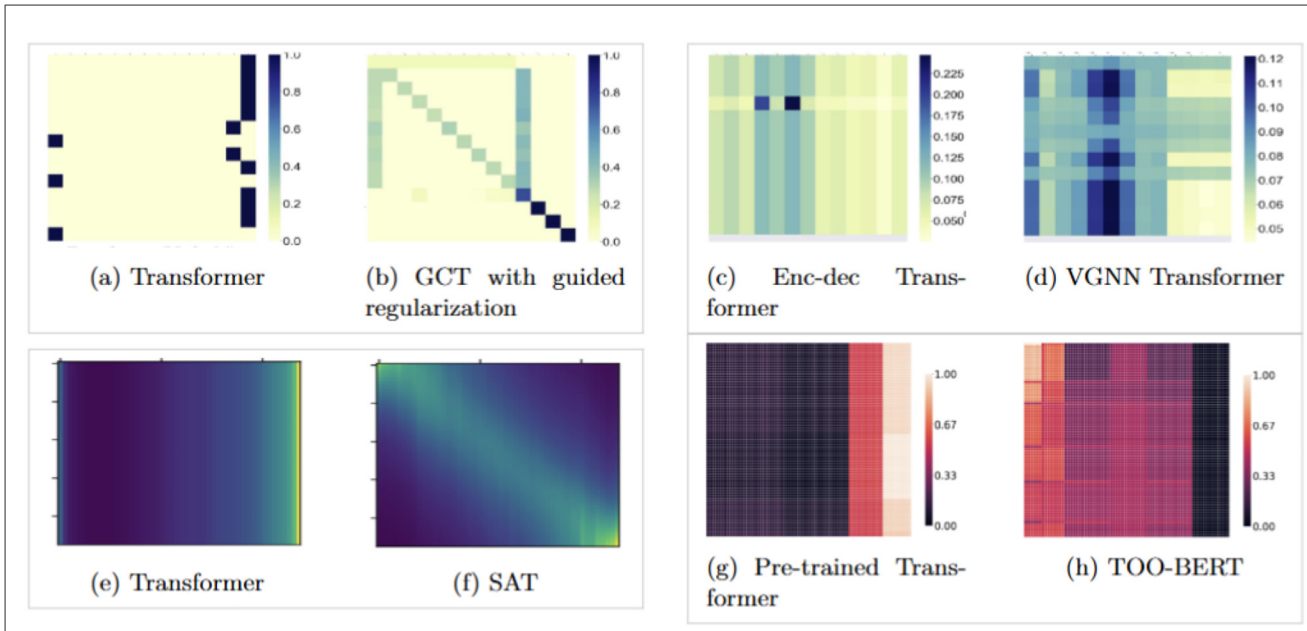

FIGURE 1
Visualization of attention score patterns for different models from previous studies and how their proposed methods helping a more complicated structure in attention scores in Transformers. **(a, b)** Transformer trained from random weights vs. Transformer trained with domain knowledge (Zhu and Razavian, 2021; Choi et al., 2020). **(c, d)** Encoder-decoder vs. VGNN using variational regularization (Zhu and Razavian, 2021). **(e, f)** Vanilla Transformer vs. SAT with temporal priors (Kim and Lee, 2024). **(g, h)** Transformer pre-trained on MLM vs. MLM with trajectory order prediction (Amirahmadi et al., 2025). **(e, f)** Had no color bars in the original papers.

3 Related works

Advanced training techniques and data augmentation have been widely adopted to improve the performance of Transformer models, especially in settings with limited labeled data. These methods aim to enhance the generalizability and robustness of learned representations.

Several methods modify self-attention to better learn intricate local and global attentions between different tokens. Hassani et al. (2023) introduced a sliding window attention mechanism to localize attention spans and improve efficiency. Ding et al. (2023) reduced attention complexity by segmenting key, query, and value inputs and sparsifying their interactions, allowing Transformers to better model both short- and long-range dependencies. Positional encoding has also been a target for improvement: Su et al. (2024) and Press et al. (2021) enhanced distant token interaction by encoding absolute positions with rotation matrices or distance-based penalties on query-key attention scores. While these methods are effective, they often require structural changes to the attention mechanism, making them less compatible with pre-trained models and harder to integrate into existing pipelines.

Data augmentation is another solution to tackle the data scarcity challenge, but it is particularly challenging in discrete domains like medical codes or text, where small changes can drastically alter semantic meaning and interpolation is not well-defined (Chen et al., 2020). To address this, researchers have proposed augmenting models during training or fine-tuning by injecting noise into internal representations (Jain et al., 2023; Zehui et al., 2019; Yuan et al., 2022; Wornow et al., 2023; Wu et al., 2023). Injecting Gaussian noise into activations has been shown to help models converge to smoother minima, improving generalization, calibration, and robustness to perturbations (Camuto et al., 2020). Zhu et al. (2019) enhanced the performance of BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) by adding adversarial noise to word embeddings, a technique later extended to graph neural networks by Kong et al. (2022) for improved out-of-distribution generalization. In the self-attention space, Zehui et al. (2019) proposed DropAttention, which randomly masks and expands attention scores to regularize focus. Similarly, Wu et al. (2023) introduced adversarial structural biases to attention matrices, though at the cost of increased training complexity.

Wornow et al. (2023) injected Gaussian noise into the latent space of an encoder-decoder model for better image captioning, while Yuan et al. (2022) perturbed hidden representations during fine-tuning to marginally improve language model performance. Most notably, Jain et al. (2023) introduced NEFTune, which adds calibrated uniform noise to embedding vectors during fine-tuning—resulting in significant improvements for models like LLaMA-1 and LLaMA-2. Inspired by these efforts, we compare our method with NEFTune and propose a new approach that directly perturbs the attention scores, encouraging the model to learn richer contextual dependencies across sequences. Here, We investigate augmenting the self-attention scores—central to modeling event dependencies—by injecting and smoothing adaptive Gaussian noise. Unlike prior methods that perturb embeddings or hidden states, our approach directly improves attention behavior without changing the model architecture, enhancing the learned representation in a lightweight and effective way.

4 Methods

4.1 Adaptive noise-augmented attention

In this subsection, we introduce, Adaptive Noise-Augmented Attention (ANAA), a simple yet effective two-step augmentation technique designed to improve the learned representations in Transformer models by directly augmenting the attention scores during fine-tuning (Algorithm 1). This method enhances attention dynamics without modifying the computational graph, making it compatible with any pre-trained Transformer encoder.

ANAA operates by first injecting adaptive Gaussian noise into the attention score matrix and then applying a smoothing operation using a Gaussian kernel. This process encourages the model to explore the attention patterns and strengthens context modeling. The augmented attention is computed as:

$$\text{ANAA} = \left((A_h + \sim \mathcal{N}(\mu, \sigma^2_{GN})) * n_{\sigma_{eh}}\right) V \tag{5}$$

Here, the Gaussian noise $\mathcal{N}(\mu, \sigma^2_{\text{GN}})$ is computed adaptivly based on the learned attention during training:

$$\mu = \frac{1}{n^2} \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} A_{i,j} \tag{6}$$

$$\sigma_{\text{GN}} = \sqrt{\frac{1}{n-1} \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} (A_{i,j} - \mu)^2} \tag{7}$$

The smoothing kernel $n_{\sigma_{\text{eh}}}[i,j]$ is a 2D Gaussian distribution:

$$n_{\sigma_{\text{eh}}}[i,j] = \frac{1}{2\pi\sigma^2_{\text{eh}}} e^{-\frac{1}{2}\left(\frac{i^2+j^2}{\sigma^2_{\text{eh}}}\right)} \tag{8}$$

Input: $D_{\texttt{fine-tuning}} = \{(X_i, y_i)\}_1^N$ tokenized dataset, embedding layer $\texttt{emb}(\cdot)$, attention score matrix A_h, normal noise $\mathcal{N}(\mu, \sigma^2_{\text{GN}})$, two-dimensional Gaussian noise $n_{\sigma_{\text{eh}}}$, rest of the model $f(\cdot)$

Parameter: Normal noise $\mu, \sigma^2_{\text{GN}}$ calculated from A_h, event horizon hyperparameter σ_{eh} based on the data charecterstic needs to adjust the smoothing noise

Initialize θ from a pre-trained model
repeat
  Sample $(X_i, y_i) \sim D_{\texttt{fine-tuning}}$
  $X_{\text{emb}} \leftarrow \texttt{emb}(X_i)$
  for each Attention Head A_h in Transformer Block **do**
    $A_h(X_{\text{attn}}) \leftarrow A_h(X_{\text{emb}}) + \mathcal{N}(\mu, \sigma^2_{\text{GN}})$
    $A_h(X_{\text{attn}}) \leftarrow \texttt{Convolve}(A_h(X_{\text{attn}}), n_{\sigma_{\text{eh}}})$
    $H_h(X_{\text{attn}}) \leftarrow A_h(X_{\text{attn}})V$
  end for
  $\texttt{MultiHead}(H) \leftarrow \texttt{concat}(H_0(X_{\text{attn}}), \ldots, H_h(X_{\text{attn}}))$
  $\hat{y}_i \leftarrow f(\texttt{MultiHead}(H))$
  $\theta \leftarrow \texttt{opt}(\theta, \texttt{loss}(\hat{y}_i, y_i))$
until Stopping criteria met or maximum iterations reached

Algorithm 1. Fine-tuning Transformer encoder with ANAA.

where σ_{eh} is a tunable hyperparameter representing the event horizon, controlling and adjusting the extent of the smoothing. The convolution operation $*$ applies this kernel over the noise-augmented attention matrix:

$$f[i,j] * n_\sigma[i,j] = \frac{1}{2\pi\sigma^2} \sum_{m=1}^{k} \sum_{n=1}^{k} e^{-\frac{1}{2}\left(\frac{m^2+n^2}{\sigma^2}\right)} f[i-m, j-n] \tag{9}$$

where $k = 2\pi\sigma$ is the kernel size.

This smoothing step modulates the added noise, reinforcing stronger attention patterns while allowing for broader exploration in attentions space. The noise parameters μ and σ_{GN} are computed independently for each attention head to preserve head-specific attention dynamics during training. Figure 2 illustrates the full ANAA mechanism.

Adding adaptive Gaussian noise $\sim \mathcal{N}(\mu, \sigma^2_{\text{GN}})$ to the attention scores helps the model escape sub-optimal solutions and promotes learning more diverse interactions between events. The subsequent Gaussian convolution adjusts the magnitude and distribution of the injected noise, encouraging the model to focus on more meaningful and effective attention patterns.

During inference, stochasticity from the added noise is removed by replacing it with its expected value μ, ensuring deterministic predictions:

$$\text{ANAA} = ((A_h + \mu) * n_{\sigma_{eh}})V \tag{10}$$

The computational complexity of ANAA is $O(n^2)$ (for more details, see the Supplementary material Section 1.9, and since it's primarily used during fine-tuning with limited labeled samples, the additional cost is negligible.

4.2 Mechanistic rationale: Why ANAA works

Figure 4 reveals that, in the absence of augmentation, many attention heads converge to a degenerate two-point distribution: each weight is either exactly 0 ("off") or 1 ("on"), with masses $1-\alpha$ and α, respectively. ANAA first perturbs the Attention scores with adaptive Gaussian noise whose variance scales as $\alpha(1-\alpha)$ (Supplementary material Section 1.5). So, every Dirac spike is broadened into a narrow normal curve, turning the rigid on/off pattern into a bimodal continuous distribution that expresses graded less-important vs. more-important scores. The subsequent Gaussian convolution (Supplementary material Section 1.6) behaves as a data-adaptive low-pass filter: it suppresses high-frequency artifacts and interpolates between neighboring tokens, so the random, isolated spikes introduced by the noise disappear.

Taken together, ANAA can be viewed as a variance-scaled, structured drop-connect regularizer (Supplementary material Section 1.7), analogous to–but more principled than–classical dropout, which disconnects token pairs with an independent Bernoulli mask. ANAA instead perturbs each attention score additively, so every mini-batch sees a different, spatially smoothed view of the inputs relations.

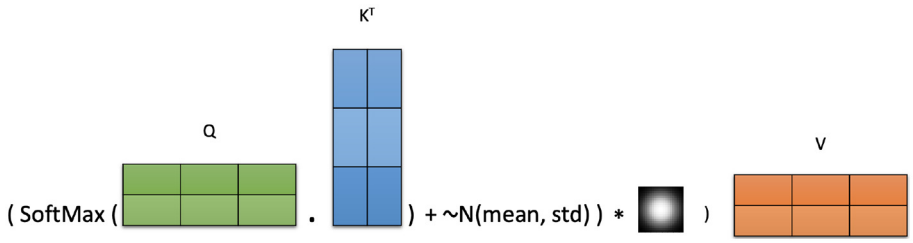

FIGURE 2
Adaptive Noise-Augmented Attention (ANAA) mechanism.

5 Experiments

5.1 Datasets

In our study, we utilized medical data from two sources: the MIMIC-IV (Johnson et al., 2020) hosp module and the Malmö Diet and Cancer Cohort (MDC) (Berglund et al., 1993) dataset, approved by the Ethics Review Board of Sweden (Dnr 2023-00503-01). Each EHR trajectory represents a sequence of temporally structured health events. The MIMIC-IV dataset includes 173,000 patient records across 407,000 visits from 2008 to 2019, with 10.6 million medical codes. The MDC dataset, from a cohort study in Sweden, comprises 30,000 individuals with 531,000 visits from 1992 to 2020, offering a more extended patient history—257 codes per patient on average, compared to MIMIC-IV's 61. To ensure consistency, we used only ICD and ATC codes, the only types available in MDC at the beginning, aligning with prior work like Med-BERT on diagnosis codes for risk prediction.

Both datasets use ICD and ATC codes for disease and medication classification. We randomly split each cohort into 70% for pre-training, 20% for fine-tuning, and 10% for testing. After preprocessing, MIMIC-IV had 2,195 unique ICD-9 and 137 ATC-5 codes, while MDC had 1,558 ICD-10 and 111 ATC-5 codes. To assess the generalizability and robustness of our results, the fine-tuning dataset was split into 5 folds. The model was fine-tuned on 4 folds with early stopping on the remaining fold, repeated 5 times with different validation sets. We reported the mean and standard deviation of the AUC on the unseen test dataset. For details, refer to the dataset availability, specifications and implementation details in the Supplementary material Sections 1.1, 1.2, 1.4.

5.2 Problem formulation

Each dataset D comprises a set of patients P, $D = \{P^1, P^2, \ldots, P^{|D|}\}$. In our study, we considered a total of $|D| = 172,980$ patients for MIMIC-IV and $|D| = 29,664$ patients for the MDC cohort. We represent each patient's longitudinal medical trajectory through a structured set of visit encounters as a sequence of events. This representation is denoted as $P^i = \{V_1^i, V_2^i, \ldots, V_O^i\}$, where O represents the total number of visit encounters for patient i. Each visit $V_j^i = I_j \cup M_j$ is the union of all diagnosis codes $I_j \subset I$ and prescribed medications $M_j \subset M$ that are recorded for the P^i at visit V_j^i. To reduce sparsity, we excluded less frequently occurring medical codes and retained only the initial 4 digits of ICD and ATC codes.

To guide the model in understanding changes in encounter times and the structure of each patient's trajectory, similar to BERT, we employed special tokens. A $[CLS]$ token is placed at the beginning of each patient's trajectory, while a $[SEP]$ token is inserted between visits. Each visit represents a set of diagnoses and medications recorded within a specific time span, and the $[SEP]$ token separates the sets of medical codes from one visit to the next. Consequently, each patient's trajectory is represented as $P^i = \{[CLS], V_1^i, [SEP], V_2^i, [SEP], \ldots, V_O^i, [SEP]\}$, providing the model with valuable context for analysis and prediction.

Here, we evaluated our models on 3 downstream tasks e_{dt} [Heart Failure (HF), Alzheimer's Disease (AD), Prolonged Length of Stay on the next visit (PLS) predictions], where the model predicts the incidence of the first HF ($I_{N=HF}$) or AD ($I_{N=AD}$) ICD codes or the presence of PLS ($PLS_N = 1$) on the N^{th} visit, given the patient's previous history of medical codes, $[V_1^i : V_{N-1}^i]$, as a sequence of temporally structured health events:

$$\mathbb{P}(e_{dt} \in V^N \mid P^i = \{[CLS], V_1^i, [SEP], V_2^i, [SEP], \ldots, V_{N-1}^i, [SEP]\}) \tag{11}$$

For each patient's trajectory, if there were no occurrences of the target events e_{dt}, it is considered a negative case; otherwise, we exclude the first visit with the target and all subsequent visits and consider it a positive case. All ATC codes related to HF treatment are excluded to avoid timing-related noise and non-trivial predictions. Initially, models exhibited bias toward longer visit histories, confounding risk predictions. To address this, we excluded trajectories with fewer than 30 visits in the MDC dataset and fewer than 10 visits in the MIMIC-IV dataset. This ensured balanced visit histories between positive and negative cases, resulting in averages of 19 visits in the MDC dataset and 9 visits in the MIMIC-IV dataset, aligning with their overall dataset averages prior to preprocessing. Table 1 summarizes the number of positive and negative cases after these preprocessing steps.

5.3 List of models

To thoroughly investigate the impact of the proposed ANAA augmentation, we compared the performance of following conventional and deep learning models on downstream tasks of HF, AD, and PLS prediction using both the MDC and MIMIC-IV

TABLE 1 Number of positive and negative samples in each downstream task.

Task	Positive	Negative
PLS prediction	2,429	6,360
HF prediction (MIMIC-IV)	243	641
AD prediction	245	2,628
HF prediction (MDC)	103	301

datasets. These models were trained either from scratch or initiated from pre-trained weights, fine-tuned on the fine-tuning dataset, and evaluated on the test dataset. We set the tunable event horizon parameter to $\sigma_{eh} = 1.0$ (kernel size = 6) for the ANAA on the MDC dataset and $\sigma_{eh} = 0.33$ (kernel size = 2) on the MIMIC IV after fine-tuning on the fine-tuning dataset. Except fir HF prediction in the MDC, different σ_{eh}, slightly changes the ANAA performance. For more details see Supplementary material Section 1.3.

5.3.1 Models with proposed RNA/ANAA

- **Transformer with ANAA**: This model incorporates ANAA into all self-attention heads of a randomly initialized Transformer.
- **Transformer pre-trained on MLM with Raw Noise injected Attention (RNA)**: In this approach, $\mathcal{N}(\mu, \sigma^2_{GN})$ (normal noise with adaptive parameters) is added to all self-attention heads of a pre-trained Transformer. This experiment allows us to isolate the impact of the noise injection from the smoothing effect of Gaussian convolution.
- **Transformer pre-trained on MLM with ANAA**: This model incorporates ANAA into all self-attention heads of the pre-trained Transformer.

Baseline model details and results are provided in Supplementary material Section 1.11.

5.4 Evaluation on downstream tasks

The results are summarized in Table 2 and suggest that adding ANAA improves the AUC of pre-trained Transformers, potentially positioning them as one of the state-of-the-art methods for outcome prediction on temporal structured health data. Specifically, on the MDC dataset, the AUC for HF and AD prediction increased to 74.5% and 73.2%, respectively, while on the MIMIC-IV dataset, the AUC for HF prediction reached 87.2%. The addition of ANAA resulted in statistically significant improvements for HF prediction on both the MDC and MIMIC-IV datasets for the MLM pre-trained Transformer. Furthermore, the improvement in AD prediction was considerable, showcasing the effectiveness of ANAA augmentation. However, incorporating ANAA did not significantly alter the performance of PLS prediction. Additionally, applying ANAA to randomly initialized Transformers boosted the AUC for PLS prediction to 60.2%, with negligible effects on other downstream tasks. To delve deeper into the impact of each noise injection and smoothing augmentation term, we solely added the normal noise to the pre-trained Transformer. This experiment revealed that the noise injection alone had a more pronounced effect on downstream tasks in the MIMIC dataset, whereas the combined (ANAA) terms exhibited greater impacts on the downstream tasks in the MDC dataset, particularly associated with its longer sequences.

5.5 Performance boost on data insufficiency

One of the advantages of using pre-trained Transformers is their robustness and performance in situations of data insufficiency, observed in both NLP (Brown et al., 2020) and temporal health data (Rasmy et al., 2021). Here, we investigated the effect of applying ANAA on model performance for HF prediction with reduced data sample sizes. We decreased the fine-tuning sample size to 50%, 20%, and 10%, respectively. The performance of the pre-trained Transformer with and without ANAA, was compared on both the MDC and MIMIC-IV datasets. Figures 3a, 4 shows that ANAA improves the model performance by around 3% in HF prediction on the MIMIC-IV dataset across all data sample sizes. Similarly, Figures 3b, 4 demonstrates that ANAA consistently outperforms the baseline in HF prediction on the MDC dataset, even with a 50% reduction in training samples. However, its superiority diminishes with less data.

5.6 VS hidden representation augmentation

We first compared ANAA with other hidden representation augmentation methods proposed for augmenting different layers of pre-trained Transformers. Specifically, we assess the impact of injecting noise into various components of the network, such as hidden layers and feedforward modules, as explored in works like HyPe (Yuan et al., 2022) and Neftune (Jain et al., 2023). Our objective is to evaluate whether augmenting self-attention scores, where contextual dependencies are explicitly encoded, is more effective than augmenting other internal representations.

As shown in Table 3, although NefTune (Jain et al., 2023) enhances the performance of pre-trained Transformers in HF prediction across both datasets, ANAA consistently outperforms both NefTune and feedforward noise augmentation in predicting outcomes. While ANAA demonstrates superior performance in this context, NefTune has the advantage of being computationally lighter. However, since both methods are applied during fine-tuning, the computational demands are not a significant concern.

5.7 VS naive masking

Randomly masking the attention score matrix during training can be seen as an extreme form of RNA augmentation. Instead of adding normal noise to perturb relationships between events in a sequence, naive masking directly disrupts these relationships by summing each element with 0 or $-A_{h_{ij}}$, effectively breaking the

TABLE 2 Average AUC values (%) and standard deviation for different methods for the HF prediction, AD prediction, and PLS prediction downstream tasks on the test datasets.

Model/downstream task	HF prediction (MDC)	AD prediction (MDC)	HF prediction (MIMIC-IV)	PLS prediction (MIMIC-IV)
Transformer	71.4 (0.5)	70.5 (0.8)	84.2 (1.4)	54.4 (0.8)
Transformer+ ANAA	72.1 (2.7)	70.4 (0.6)	83.2 (2.5)	60.2 (1.2)
Transformer pre-trained on MLM	72.2 (2.5)	72.2 (1.1)	85.2 (1.1)	60.3 (1.3)
Transformer pre-trained on MLM+ RNA	72.6 (1.9)	71.4 (1.0)	86.5 (1.2)	**60.7 (0.6)**
Transformer pre-trained on MLM+ ANAA	**74.5 (2.9)**	**73.2 (0.3)**	**87.2 (0.4)**	60.3 (0.7)

Boldface indicates the best-performing model.

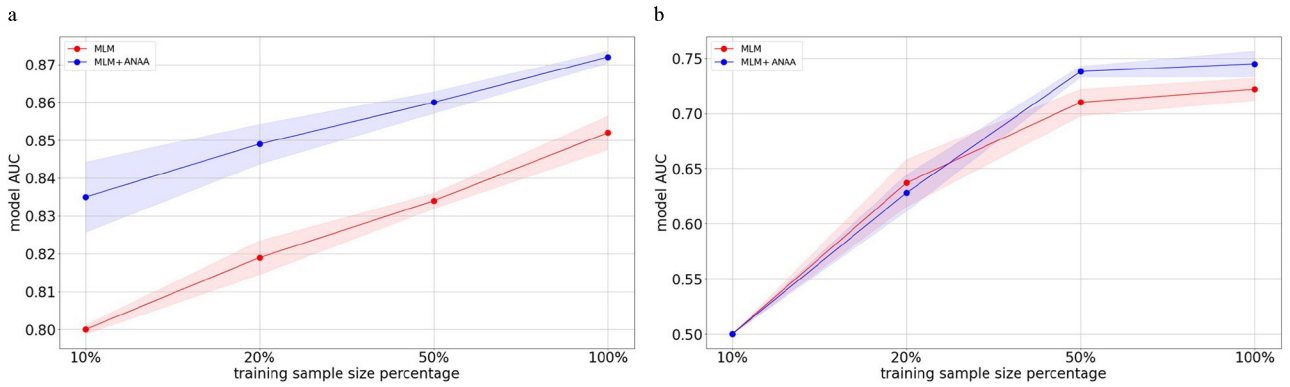

FIGURE 3
Impact of ANAA on AUC for HF prediction across different fine-tuning sample sizes in the MIMIC-IV and MDC datasets. The red line shows the AUC of a Transformer model pre-trained on MLM without augmentation; the blue line shows the AUC of the same model augmented with ANAA. In MIMIC-IV, MLM+ANAA consistently outperforms the MLM baseline at all sample sizes. In MDC, MLM+ANAA outperforms the baseline up to the 50% training size; at smaller sizes, its performance converges to that of the baseline due to the limited number of HF-positive samples in the MDC dataset. **(a)** AUC values for HF prediction across fine-tuning sample sizes on the MIMIC-IV test set. **(b)** AUC values for HF prediction across fine-tuning sample sizes on the MDC test set.

connections between tokens. We compared our method with naive self-attention masking, as described by Wu et al. (2023), which introduces a bias in the structure of self-attentions:

$$A_h = \text{softmax}\left(\frac{Q_h K_h^T}{\sqrt{d_k}} + M\right), \quad M \in \{0, -\infty\}^{N\times N}, \qquad (12)$$

where $M_{i,j} = -\infty$ with $p = 0.2$, optimized based on performance on the fine-tuning dataset. We extended it to DropAttention (Zehui et al., 2019), which expands the mask with a span length ω and we set $\omega =$ Kernel size. However, neither naive masking nor DropAttention improved the performance of the pre-trained Transformer for HF prediction on the MDC and MIMIC-IV datasets. Instead, these methods only increased the number of training iterations required for convergence (see Table 3). While these techniques can help mitigate overfitting, their overly aggressive regularization often disrupts critical dependencies within sequences, leading to unstable training and poorer overall performance, especially on complex healthcare prediction tasks. In contrast, ANAA introduces controlled perturbations that balance the attention distribution and prevent over-reliance on specific patterns, thereby preserving essential relationships in the data and promoting more robust and effective representations (see Supplementary material Section 1.7 for a justification of ANAA as a structured variant of dropout).

5.8 Effect of ANAA on self-attention behavior

Analyzing self-attention weights and attention score matrices can highlight how Transformers prioritize relationships between events, shedding light on their internal logic and behavior (Clark et al., 2019; Kovaleva et al., 2019; Hao et al., 2021). To assess the effect of ANAA and compare it with normal noise injection (RNA), we analyzed attention score distributions in models fine-tuned on all downstream tasks.

We plotted histograms of attention scores across all heads and samples from the test split, scaling each head's scores to the [0, 1] range (Figure 4). In the bottom row of the figure, we observe that attention scores from the fine-tuned vanilla Transformer tend to cluster near 0 or 1, forming a near-binary (binomial-like) distribution. This pattern suggests overconfidence and limited exploration of dependencies across tokens.

In contrast, the middle row shows that RNA—injecting Gaussian noise during training—broadens the

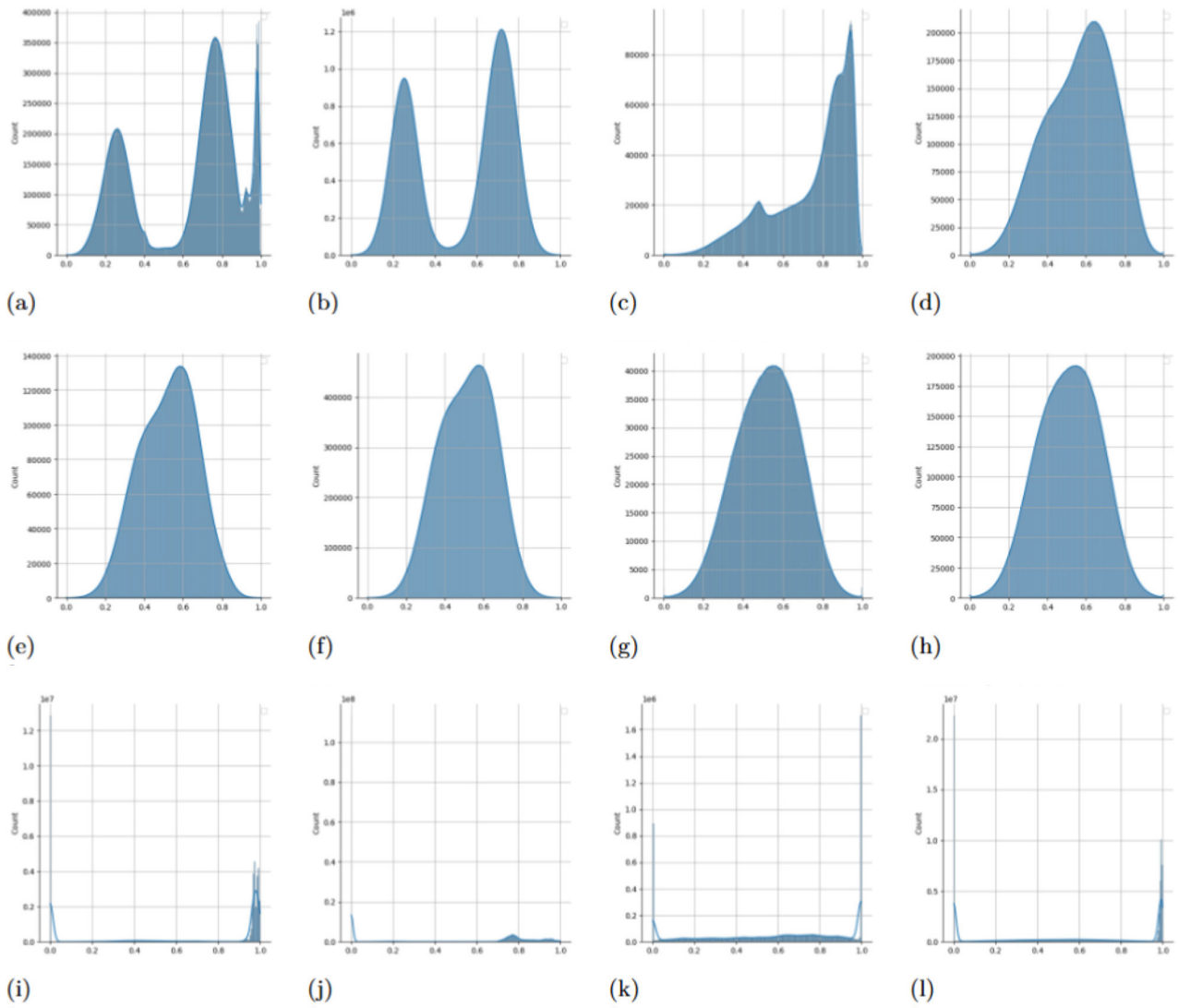

FIGURE 4
Comparison of the impact of ANAA on self-attention score distributions in fine-tuned models. Attention scores from each head are individually scaled to the [0, 1] range before plotting their distributions. **(a)** Pre-trained Transformer + Smoothed noise. **(b)** Pre-trained Transformer + ANAA. **(c)** Pre-trained Transformer + ANAA. **(d)** Pre-trained Transformer + ANAA. **(e)** Pre-trained Transformer + RNA. **(f)** Pre-trained Transformer + RNA. **(g)** Pre-trained Transformer + RNA. **(h)** Pre-trained Transformer + RNA. **(i)** Pre-trained Transformer. **(j)** Pre-trained Transformer. **(k)** Pre-trained Transformer. **(l)** Pre-trained Transformer.

distribution, encouraging attention heads to explore more diverse and weaker connections. This leads to overlapping attention patterns and increased representation diversity. A mathematical explanation for this phenomenon is provided in Supplementary material Section 1.5.

The top row demonstrates the effect of ANAA, which combines noise injection with Gaussian smoothing. This operation retains the diversity introduced by noise while stabilizing the attention pattern, restoring smoother and more informative distributions. The smoothing step dampens extreme noise while allowing the model to refine its exploration of differnt interactions.

To further investigate, we visualized the attention score matrices from models fine-tuned on a representative test sample from the HF prediction task on the MDC dataset (Figure 5). Comparing the original and smoothed attention scores, we observe that ANAA promotes broader attention coverage, with activation scores scaled to the [0, 1] range. Figure 5 illustrates an attention head from the first layer, confirming that ANAA leads to more distributed attention patterns. Additional examples from the MIMIC-IV dataset are provided in the Supplementary material Section 1.10.

However, it is important to note that the heat-maps in Figures 4, 5 are intended as qualitative diagnostics of how ANAA redistributes attention–not to explain the model's decisions. As shown in prior work, attention weights can often be manipulated without affecting model outputs, meaning they are not a reliable source of explanation (Hao et al., 2021; Jain and Wallace, 2019; Serrano and Smith, 2019). We therefore interpret these visualizations only as evidence that ANAA breaks the near-binary pattern observed in the baseline model; attributing clinical

TABLE 3 Comparing ANAA with naive masking and other hidden representation augmentation methods.

Model/downstream task	HF prediction (MDC)	HF prediction (MIMIC-IV)
Transformer pre-trained on MLM	72.2 (2.5)	85.2 (1.1)
Transformer pre-trained on MLM+ Naive masking	70.00 (1.5)	85.1 (0.7)
Transformer pre-trained on MLM+ DropAttention	69.7 (1.1)	84.9 (1.3)
Transformer pre-trained on MLM+ NEFTune ($\alpha = 5$)	73.6 (3.2)	85.2 (0.7)
Transformer pre-trained on MLM+ NEFTune ($\alpha = 10$)	73.1 (1.7)	85.5 (0.4)
Transformer pre-trained on MLM+ noise in the feedforward ($\alpha = 5$)	73.7 (2.2)	85.0 (1.2)
Transformer pre-trained on MLM+ noise in the feedforward ($\alpha = 10$)	72.5 (4.4)	84.5 (0.8)
Transformer pre-trained on MLM+ ANAA	**74.5 (2.9)**	**87.2 (0.4)**

The table shows the average AUC values (%) and standard deviation across HF prediction tasks on the MDC and MIMIC-IV datasets. Boldface indicates the best-performing model.

relevance to specific codes and specific codes with each other in this context would require dedicated methods such as Integrated Gradients (Sundararajan et al., 2017) and can be investigated further in future work.

5.8.1 Effect of ANAA on the receptive field

The self-attention mechanism is designed to capture both long and short-range dependencies effectively. To quantitatively assess the impact of RNA and ANAA on the receptive field, we plot the median values of attention score matrix A_h for each event with respect to all previous and subsequent events ($i - j, A_{h_{i,j}}$) -i, j are positions of e_i, e_j in the sequence of events-across all test samples for HF and AD predictions on the MDC (Figure 6). Transformers pre-trained on MLM typically allocate more attention weight to recent events, often in a monotonous fashion. Incorporating RNA regularization reduces the steepness of this attention distribution, allowing events to receive more balanced attention, not solely based on their proximity to recent events. Ultimately, applying ANAA, preserves the benefits of RNA by providing a more equal distribution of attention within a local neighborhood, while simultaneously reducing the emphasis on very distant past events.

6 Discussion

This study demonstrates that ANAA—a simple two-step fine-tuning augmentation—consistently enhances the discriminative performance of pre-trained Transformers on longitudinal EHR data, without altering their architecture. Compared to the hidden representation augmentation and a range of established regularizers, it yields superior results over vanilla fine-tuning. ANAA produced consistent AUC gains on HF and AD prediction tasks in two different EHR corpora (MDC and MIMIC-IV). On HF prediction, for example, the MLM-pre-trained baseline rose from 72.2 to 74.5 AUC on MDC and from 85.2 to 87.2 AUC on MIMIC-IV after applying ANAA. These gains persisted even under label-scarce conditions, maintaining ∼3 percentage-point improvements. These findings suggest that judicious noise injection at the level of self-attention—followed by controlled Gaussian smoothing—can encourage pre-trained transformers to explore and learn more robust, generalizable patterns.

We further investigated that Transformers pre-trained via MLM—while typically outperforming models without pre-training—can exhibit overconfident, sparse attention patterns during fine-tuning. Attention histograms reveal that conventional fine-tuning drives many heads toward almost binary (0/1) weights, indicating over-confident, brittle dependencies. ANAA counteracts this by injecting adaptive Gaussian noise, which broadens the attention distribution and encourages heads to sample a richer set of relational cues. The subsequent smoothing step restores coherent structure. As shown analytically in Supplementary material Section 1.5, this mechanism effectively acts as a variance-scaled, shifting the attention score distribution from deterministic and binary to probabilistic and continuous, to explore alternative dependencies.

Compared to other augmentation methods such as NEFTune (Jain et al., 2023) and HyPe (Yuan et al., 2022)—which add noise in the embedding or feed-forward layers—ANAA achieves larger and more consistent performance gains. In contrast, naive attention masking or DropAttention (Wu et al., 2023; Zehui et al., 2019) degraded results. This highlights the importance of *where* noise is injected: perturbing the self-attention scores—the core mechanism for modeling token interactions—yields greater benefit than altering downstream representations.

While ANAA consistently improves performance across the two studied EHR datasets, several caveats remain. First, all experiments were conducted on structured, diagnosis- and medication-coded timelines (MIMIC-IV and MDC); Although our experiments focus on a standard Transformer encoder for clarity and control, ANAA is modular by design and can be integrated into other clinical Transformer models such as BEHRT (Li et al., 2020), Med-BERT (Rasmy et al., 2021), or Hi-BEHRT (Li et al., 2022); exploring such integrations is a promising direction for future work. More broadly, how well ANAA generalizes to other data modalities –such as free text, imaging, or genomics –and to models pre-trained with alternative objectives such as contrastive learning (e.g., BYOL; Grill et al., 2020) also remains to be explored. Second, ANAA introduces additional hyperparameters. Although the sensitivity analysis in Supplementary Table S3 suggests the method is robust across a range of values, some tuning is still required. Third, the computational overhead introduced by noise injection and smoothing increases both memory usage and training time, which may become a limitation for very long sequences or resource-constrained environments. Fourth, in settings with extremely low data regimes or highly unbalanced labels, ANAA's implicit Augmentation provides some benefit but is not sufficient on its own. Finally, the effect of model augmentations, like ANAA, on model interpretability warrants further study, particularly in safety-critical applications.

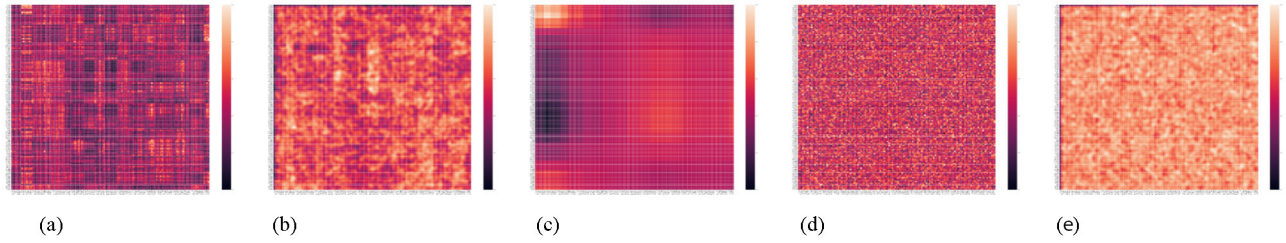

FIGURE 5
Comparing the impact of ANAA on the self-attention score weights for five fine-tuned models on HF prediction on the MDC dataset for a specific test sample. Here, the attention scores are scaled within 0 and 1. **(a)** Transformer. **(b)** Transformer + ANAA. **(c)** Pre-trained Transformer. **(d)** Pre-trained Transformer + RNA. **(e)** Pre-trained Transformer + ANAA.

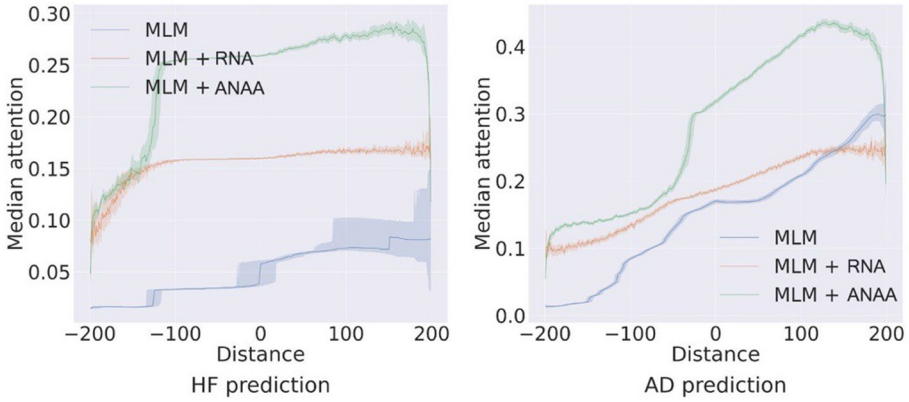

FIGURE 6
Impact of ANAA on the receptive field of the self-attentions for HF and AD prediction on the MDC dataset.

7 Conclusion

We introduced *Adaptive Noise-Augmented Attention* (ANAA), a lightweight and effective method for enhancing the fine-tuning of pre-trained Transformers. ANAA directly augments the self-attention scores with adaptive Gaussian noise and applies a smoothing convolution using a Gaussian kernel, encouraging the model to explore more diverse attention patterns while preserving critical dependencies.

We demonstrated that pre-trained Transformers, when fine-tuned on limited EHR datasets, often converge to overly sharp attention distributions—overfitting to local patterns and failing to capture broader contextual relationships. ANAA mitigates this by encouraging more diverse and stable attention distributions, leading to better generalization across tasks and data regimes. Extensive experiments on multiple clinical prediction tasks showed that ANAA consistently outperforms conventional regularization and hidden augmentation techniques.

ANAA offers a plug-and-play augmentation mechanism that operates entirely within the attention computation, requiring no modification to the model architecture or computational graph. This makes it particularly suitable for integration with existing pre-trained models.

Data availability statement

The MIMIC-IV dataset is publicly available from the PhysioNet repository [https://physionet.org/content/mimiciv/2.2/]. The Malmo Diet and Cancer Cohort data that support the findings of this study are not publicly available due to data access restrictions imposed by the Malmo Population-Based Cohorts Joint Database. However, the data are available from the corresponding author upon reasonable request and with permission from the Malmo Population-Based Cohorts Joint Database [https://www.malmo-kohorter.lu.se/malmo-cohorts].

Ethics statement

The use of the MDC dataset for this study was approved by the Ethics Review Board of Sweden (Dnr 2023-00503-01). Regarding the MIMIC-IV dataset, all protected health information (PHI) is officially deidentified. It means that the deletion of PHI from structured data sources (e.g., database fields that provide age, genotypic information, and past and current diagnosis and treatment categories) is performed in compliance with the HIPAA (Health Insurance Portability and Accountability Act) standards in order to facilitate public access to the datasets.

Author contributions

AA: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. FE: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Resources, Supervision, Validation, Visualization, Writing – review & editing. MO: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This study was conducted as part of the AIR Lund (Artificially Intelligent use of Registers at Lund University) research environment and was funded by the Swedish Research Council (VR, grant 2019-00198). Additional support was provided by CAISR Health, funded by the Knowledge Foundation (KK-stiftelsen) in Sweden (grant 20200208 01 H).

Acknowledgments

We thank Jonas Björk and Olle Melander for facilitating access to the data and for their valuable guidance in understanding and interpreting the dataset.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2025.1663484/full#supplementary-material

References

Amirahmadi, A., Etminani, F., Björk, J., Melander, O., and Ohlsson, M. (2025). Trajectory-ordered objectives for self-supervised representation learning of temporal healthcare data using transformers: Model development and evaluation study. *JMIR Med. Inform*. 13:e68138. doi: 10.2196/68138

Amirahmadi, A., Ohlsson, M., and Etminani, K. (2023). Deep learning prediction models based on ehr trajectories: a systematic review. *J. Biomed. Inform*. 144:104430. doi: 10.1016/j.jbi.2023.104430

Amos, I., Berant, J., and Gupta, A. (2023). Never train from scratch: fair comparison of long-sequence models requires data-driven priors. *arXiv preprint arXiv:2310.02980*. doi: 10.48550/arXiv.2310.02980

Berglund, G., Elmståhl, S., Janzon, L., and Larsson, S. (1993). The malmo diet and cancer study. Design and feasibility. *J. Intern. Med*. 233, 45–51. doi: 10.1111/j.1365-2796.1993.tb00647.x

Boll, H. O., Amirahmadi, A., Ghazani, M. M., de Morais, W. O., de Freitas, E. P., Soliman, A., et al. (2024). Graph neural networks for clinical risk prediction based on electronic health records: a survey. *J. Biomed. Inform*. 151:104616. doi: 10.1016/j.jbi.2024.104616

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*. doi: 10.48550/arXiv.2108.07258

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Adv. Neural Inf. Process. Syst*. 33, 1877–1901. doi: 10.48550/arXiv.2005.14165

Camuto, A., Willetts, M., Simsekli, U., Roberts, S. J., and Holmes, C. C. (2020). Explicit regularisation in gaussian noise injections. *Adv. Neural Inf. Process. Syst*. 33, 16603–16614. doi: 10.48550/arXiv.2007.07368

Chen, J., Yang, Z., and Yang, D. (2020). Mixtext: linguistically-informed interpolation of hidden space for semi-supervised text classification. *arXiv preprint arXiv:2004.12239*. doi: 10.18653/v1/2020.acl-main.194

Chen, X., Xie, S., and He, K. (2021). "An empirical study of training self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Los Alamitos, CA: IEEE Computer Society (Conference Publishing Services)), 9640–9649. doi: 10.1109/ICCV48922.2021.00950

Choi, E., Xu, Z., Li, Y., Dusenberry, M., Flores, G., Xue, E., et al. (2020). Learning the graphical structure of electronic health records with graph convolutional transformer. *Proc. AAAI Conf. Artif. Intell*. 34, 606–613. doi: 10.1609/aaai.v34i01.5400

Clark, K., Khandelwal, U., Levy, O., and Manning, C. D. (2019). What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*. doi: 10.18653/v1/W19-4828

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. doi: 10.48550/arXiv.1810.04805

Ding, J., Ma, S., Dong, L., Zhang, X., Huang, S., Wang, W., et al. (2023). Longnet: scaling transformers to 1,000,000,000 tokens. *arXiv preprint arXiv:2307.02486*. doi: 10.14218/JCTH.2022.00006S

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16 × 16 words: transformers for image recognition at scale. *arXiv preprint arXiv*:2010.11929. doi: 10.48550/arXiv.2010.11929

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., et al. (2020). Bootstrap your own latent-a new approach to self-supervised learning. *Adv. Neural Inf. Process. Syst*. 33, 21271–21284. doi: 10.48550/arXiv.2006.07733

Guo, L. L., Steinberg, E., Fleming, S. L., Posada, J., Lemmon, J., Pfohl, S. R., et al. (2023). Ehr foundation models improve robustness in the presence of temporal distribution shift. *Sci. Rep.* 13:3767. doi: 10.1038/s41598-023-30820-8

Hao, Y., Dong, L., Wei, F., and Xu, K. (2021). Self-attention attribution: interpreting information interactions inside transformer. *Proc. AAAI Conf. Artif. Intell.* 35, 12963–12971. doi: 10.1609/aaai.v35i14.17533

Hassani, A., Walton, S., Li, J., Li, S., and Shi, H. (2023). "Neighborhood attention transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Los Alamitos, CA: IEEE Computer Society (Conference Publishing Services)), 6185–6194. doi: 10.1109/CVPR52729.2023.00599

Hassani, A., Walton, S., Shah, N., Abuduweili, A., Li, J., and Shi, H. (2021). Escaping the big data paradigm with compact transformers. *arXiv preprint arXiv:2104.05704*. doi: 10.48550/arXiv.2104.05704

Jain, N., Chiang, P.-y., Wen, Y., Kirchenbauer, J., Chu, H.-M., Somepalli, G., et al. (2023). Neftune: Noisy embeddings improve instruction finetuning. *arXiv preprint arXiv:2310.05914*. doi: 10.48550/arXiv.2310.05914

Jain, S., and Wallace, B. C. (2019). Attention is not explanation. *arXiv preprint arXiv:1902.10186*. doi: 10.48550/arXiv.1902.10186

Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., and Mark, R. (2020). *Mimic-iv. PhysioNet*. Available online at: https://physionet.org/content/mimiciv/1.0/ (Accessed August 23, 2021).

Kim, K. G., and Lee, B. T. (2024). Self-attention with temporal prior: can we learn more from the arrow of time? *Front. Artif. Intell.* 7:1397298. doi: 10.3389/frai.2024.1397298

Kong, K., Li, G., Ding, M., Wu, Z., Zhu, C., Ghanem, B., et al. (2022). "Robust optimization as data augmentation for large-scale graphs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Los Alamitos, CA: IEEE Computer Society (Conference Publishing Services)), 60–69. doi: 10.1109/CVPR52688.2022.00016

Kovaleva, O., Romanov, A., Rogers, A., and Rumshisky, A. (2019). Revealing the dark secrets of bert. *arXiv preprint arXiv*:1908.08593. doi: 10.18653/v1/D19-1445

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). Albert: a lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*. doi: 10.48550/arXiv.1909.11942

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., et al. (2019). Bart: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*. doi: 10.18653/v1/2020.acl-main.703

Li, L., Zhou, J., Gao, Z., Hua, W., Fan, L., Yu, H., et al. (2024). A scoping review of using large language models (LLMS) to investigate electronic health records (EHRS). *arXiv preprint arXiv*:2405.03066. doi: 10.48550/arXiv.2405.03066

Li, Y., Mamouei, M., Salimi-Khorshidi, G., Rao, S., Hassaine, A., Canoy, D., et al. (2022). Hi-behrt: hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records. *IEEE J. Biomed. Health Inform.* 27, 1106–1117. doi: 10.1109/JBHI.2022.3224727

Li, Y., Rao, S., Solares, J. R. A., Hassaine, A., Ramakrishnan, R., Canoy, D., et al. (2020). Behrt: transformer for electronic health records. *Sci. Rep.* 10:7155. doi: 10.1038/s41598-020-62922-y

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023). Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* 55, 1–35. doi: 10.1145/3560815

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). Roberta: a robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*. doi: 10.48550/arXiv.1907.11692

Pang, C., Jiang, X., Kalluri, K. S., Spotnitz, M., Chen, R., Perotte, A., et al. (2021). "Cehr-bert: incorporating temporal information from structured ehr data to improve prediction tasks," in *Machine Learning for Health* (PMLR), 239–260.

Press, O., Smith, N. A., and Lewis, M. (2021). Train short, test long: attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*. doi: 10.48550/arXiv.2108.12409

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). *Language Models Are Unsupervised Multitask Learners*. Technical Report. OpenAI. Available online at: https://openai.com/index/better-language-models/ (Accessed September 04, 2025).

Rasmy, L., Xiang, Y., Xie, Z., Tao, C., and Zhi, D. (2021). Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit. Med.* 4:86. doi: 10.1038/s41746-021-00455-y

Ren, H., Wang, J., Zhao, W. X., and Wu, N. (2021). "Rapt: pre-training of time-aware transformer for learning robust healthcare representation," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (New York, NY: Association for Computing Machinery (ACM)), 3503–3511. doi: 10.1145/3447548.3467069

Serrano, S., and Smith, N. A. (2019). Is attention interpretable? *arXiv preprint arXiv:1906.03731*. doi: 10.48550/arXiv.1906.03731

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. (2024). Roformer: enhanced transformer with rotary position embedding. *Neurocomputing* 568:127063. doi: 10.1016/j.neucom.2023.127063

Sundararajan, M., Taly, A., and Yan, Q. (2017). "Axiomatic attribution for deep networks," in *International Conference on Machine Learning* (Sydney, NSW: PMLR), 3319–3328.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. (2021). "Training data-efficient image transformers and distillation through attention," in *International Conference on Machine Learning* (Vienna: PMLR), 10347–10357.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Adv. Neural Inf. Process. Syst.* 30, 5999–6009. doi: 10.48550/arXiv.1706.03762

Wornow, M., Xu, Y., Thapa, R., Patel, B., Steinberg, E., Fleming, S., et al. (2023). The shaky foundations of large language models and foundation models for electronic health records. *NPJ Digit. Med.* 6:135. doi: 10.1038/s41746-023-00879-8

Wu, H., Ding, R., Zhao, H., Xie, P., Huang, F., and Zhang, M. (2023). Adversarial self-attention for language understanding. *Proc. AAAI Conf. Artif. Intell.* 37, 13727–13735. doi: 10.1609/aaai.v37i11.26608

Xiao, C., Choi, E., and Sun, J. (2018). Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *J. Am. Med. Inform. Assoc.* 25, 1419–1428. doi: 10.1093/jamia/ocy068

Yuan, H., Yuan, Z., Tan, C., Huang, F., and Huang, S. (2022). Hype: better pre-trained language model fine-tuning with hidden representation perturbation. *arXiv preprint arXiv:2212.08853*. doi: 10.18653/v1/2023.acl-long.182

Yuanyuan, Z., Adel, B., Mina, B., Jamil, Z., Hugues, T., Lydie, B., et al. (2025). A scoping review of self-supervised representation learning for clinical decision making using ehr categorical data. *NPJ Digit. Med.* 8, 1–15. doi: 10.1038/s41746-025-01692-1

Zehui, L., Liu, P., Huang, L., Chen, J., Qiu, X., and Huang, X. (2019). Dropattention: a regularization method for fully-connected self-attention networks. *arXiv preprint arXiv*:1907.11065. doi: 10.48550/arXiv.1907.11065

Zhu, C., Cheng, Y., Gan, Z., Sun, S., Goldstein, T., and Liu, J. (2019). Freelb: enhanced adversarial training for natural language understanding. *arXiv preprint arXiv:1909.11764*. doi: 10.48550/arXiv.1909.11764

Zhu, W., and Razavian, N. (2021). "Variationally regularized graph-based representation learning for electronic health records," in *Proceedings of the Conference on Health, Inference, and Learning* (New York, NY: Association for Computing Machinery (ACM)), 1–13. doi: 10.1145/3450439.3451855