# DIFFERENTIAL TRANSFORMER

**Tianzhu Ye**[*†‡] **Li Dong**[*†] **Yuqing Xia**[*†] **Yutao Sun**[*†‡]
**Yi Zhu**[†] **Gao Huang**[‡] **Furu Wei**[†◇]
[†] Microsoft Research [‡] Tsinghua University
https://aka.ms/GeneralAI

## Abstract

Transformer tends to overallocate attention to irrelevant context. In this work, we introduce DIFF Transformer, which amplifies attention to the relevant context while canceling noise. Specifically, the differential attention mechanism calculates attention scores as the difference between two separate softmax attention maps. The subtraction cancels noise, promoting the emergence of sparse attention patterns. Experimental results on language modeling show that DIFF Transformer outperforms Transformer in various settings of scaling up model size and training tokens. More intriguingly, it offers notable advantages in practical applications, such as long-context modeling, key information retrieval, hallucination mitigation, in-context learning, and reduction of activation outliers. By being less distracted by irrelevant context, DIFF Transformer can mitigate hallucination in question answering and text summarization. For in-context learning, DIFF Transformer not only enhances accuracy but is also more robust to order permutation, which was considered as a chronic robustness issue. The results position DIFF Transformer as a highly effective and promising architecture to advance large language models.

## 1 Introduction

Transformer [41] has garnered significant research interest in recent years, with the decoder-only Transformer emerging as the de facto standard for large language models (LLMs). At the heart of Transformer is the attention mechanism, which employs the softmax function to weigh the importance of various tokens in a sequence. However, recent studies [17, 23] show that LLMs face challenges in accurately retrieving key information from context.

As illustrated on the left side of Figure 1, we visualize the normalized attention scores assigned to different parts of the context by a Transformer. The task is to retrieve an answer embedded in the middle of a pile of documents. The visualization reveals that Transformer tends to allocate only a small proportion of attention scores to the correct answer, while disproportionately focusing on irrelevant context. The experiments in Section 3 further substantiate that Transformers struggle with such capabilities. The issue arises from non-negligible attention scores assigned to irrelevant context, which ultimately drowns out the correct answer. We term these extraneous scores as *attention noise*.

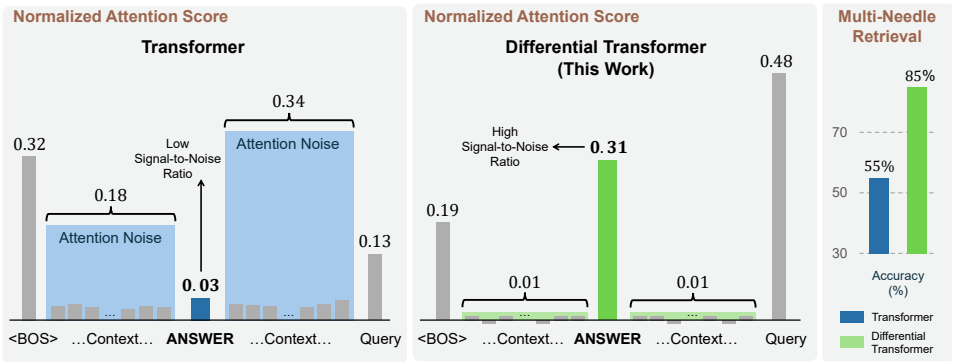


Figure 1: Transformer often over-attends to irrelevant context (i.e., attention noise). DIFF Transformer amplifies attention to answer spans and cancels noise, enhancing the capability of context modeling.

[*] Equal contribution. ◇ Corresponding author.

In this paper, we introduce Differential Transformer (a.k.a. DIFF Transformer), a foundation architecture for large language models. The differential attention mechanism is proposed to cancel attention noise with differential denoising. Specifically, we partition the query and key vectors into two groups and compute two separate $\mathrm{softmax}$ attention maps. Then the result of subtracting these two maps is regarded as attention scores. The differential attention mechanism eliminates attention noise, encouraging models to focus on critical information. The approach is analogous to noise-canceling headphones and differential amplifiers [19] in electrical engineering, where the difference between two signals cancels out common-mode noise. In the middle of Figure 1, we also present the normalized distribution of attention scores for DIFF Transformer. We observe that DIFF Transformer assigns significantly higher scores to the correct answer and much lower scores to irrelevant context compared to Transformer. The right side of Figure 1 shows that the proposed method achieves notable improvements in retrieval capability.

We conduct extensive experiments on language modeling. We scale up DIFF Transformer in terms of parameter count, training tokens, and context length. The scaling curves indicate that DIFF Transformer requires only about 65% of model size or training tokens needed by Transformer to achieve comparable language modeling performance. Moreover, DIFF Transformer outperforms Transformer in various downstream tasks. The long-sequence evaluation also shows that DIFF Transformer is highly effective in utilizing the increasing context. In addition, the experimental results demonstrate that DIFF Transformer has intriguing advantages for large language models. For example, the proposed method substantially outperforms Transformer in key information retrieval, hallucination mitigation, and in-context learning. DIFF Transformer also reduces outliers in model activations, which provides new opportunities for quantization. The findings establish DIFF Transformer as an effective and distinctive foundation architecture for large language models.

## 2 Differential Transformer

We propose Differential Transformer (a.k.a. DIFF Transformer) as a foundation architecture for sequence modeling, such as large language models (LLMs). We take a decoder-only model as an example to describe the architecture. The model is stacked with $L$ DIFF Transformer layers. Given an input sequence $x = x_1 \cdots x_N$, we pack the input embeddings into $X^0 = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N] \in \mathbb{R}^{N \times d_{\text{model}}}$, where $d_{\text{model}}$ represents the hidden dimension of the model. The input is further contextualized to obtain the output $X^L$, i.e., $X^l = \mathrm{Decoder}(X^{l-1}),\ l \in [1, L]$. Each layer consists of two modules: a differential attention module followed by a feed-forward network module. Compared to Transformer [41], the main difference is the replacement of conventional $\mathrm{softmax}$ attention with differential attention while the macro layout is kept the same. We also adopt pre-RMSNorm [46] and SwiGLU [35, 29] as improvements following LLaMA [38].

### 2.1 Differential Attention

The differential attention mechanism maps query, key, and value vectors to outputs. We use query and key vectors to compute attention scores, and then compute a weighted sum of value vectors. The critical design is that we use a pair of $\mathrm{softmax}$ functions to cancel the noise of attention scores. Specifically, given input $X \in \mathbb{R}^{N \times d_{\text{model}}}$, we first project them to query, key, and value $Q_1, Q_2, K_1, K_2 \in \mathbb{R}^{N \times d},\ V \in \mathbb{R}^{N \times 2d}$. Then the differential attention operator $\mathrm{DiffAttn}(\cdot)$ computes outputs via:

$$
\begin{aligned}
[Q_1; Q_2] = XW^Q, \quad [K_1; K_2] = XW^K, \quad V = XW^V \\
\mathrm{DiffAttn}(X) = (\mathrm{softmax}(\frac{Q_1 K_1^T}{\sqrt{d}}) - \lambda\, \mathrm{softmax}(\frac{Q_2 K_2^T}{\sqrt{d}}))V
\end{aligned} \tag{1}
$$

where $W^Q, W^K, W^V \in \mathbb{R}^{d_{\text{model}} \times 2d}$ are parameters, and $\lambda$ is a learnable scalar. In order to synchronize the learning dynamics, we re-parameterize the scalar $\lambda$ as:

$$
\lambda = \exp(\lambda_{\mathbf{q_1}} \cdot \lambda_{\mathbf{k_1}}) - \exp(\lambda_{\mathbf{q_2}} \cdot \lambda_{\mathbf{k_2}}) + \lambda_{\text{init}} \tag{2}
$$

where $\lambda_{\mathbf{q_1}}, \lambda_{\mathbf{k_1}}, \lambda_{\mathbf{q_2}}, \lambda_{\mathbf{k_2}} \in \mathbb{R}^d$ are learnable vectors, and $\lambda_{\text{init}} \in (0, 1)$ is a constant used for the initialization of $\lambda$. We empirically find that the setting $\lambda_{\text{init}} = 0.8 - 0.6 \times \exp(-0.3 \cdot (l-1))$ works well in practice, where $l \in [1, L]$ represents layer index. It is used as the default strategy in our