# Toward Efficient Sequence Modeling with Temperature assisted Information Gating
# Novel Alternative to Self Attention

M. Hamza Alvi*

May 2025

**Abstract**

Self-attention mechanisms, popularized by the Transformer architecture [2], have significantly advanced the field of deep learning. Despite their strengths in modeling global dependencies, they face challenges such as high computational costs and inefficiencies in processing long sequences. This paper offers a critical evaluation of self-attention and introduces a novel approach — **[Temperature assisted information Gating]** to surpass Self Attention efficiently and overcoming these issues.

## 1 Limitations of Self-Attention Mechanisms

Self-attention operates by computing relationships between all pairs of input tokens:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where $Q$, $K$, and $V$ represent the queries, keys, and values, and $d_k$ is the dimension of the key vectors.

While this formulation enables context-aware representations, it incurs **quadratic time and space complexity** in input length [2]. Linformer [3] addresses this by projecting $K$ and $V$ to lower dimensions, reducing complexity to linear. Performer [1] introduces kernelized attention to achieve similar efficiency.

However, these solutions come with trade-offs in expressiveness, and the underlying interpretability and inductive bias issues remain.

## 2 Dot Product: Rethinking After-effects of dot Product

The core idea is to think what the effect of dot product on the values of hidden states matrix. If the numbers are smaller suppose 0.1 x 0.1 instead of increasing it decrease and becomes 0.001 while numbers greater than 1 usually increase its magnitude greatly. What does it mean, how we even interpret it, So if the model penalized some weights to be small it becomes even more smaller effectively supressing and on other hand if it encourages some weights to increase the effect becomes more magnified. It basically acts as an information gating system which the model learns its way during training.

*Email: mhamzaawan7786@gmail.com. GitHub Repo: https://github.com/Frankenstein-hsa/Pelican

## 3 Effect of Temperature on Probabilities

The idea is that can we do it without expensive dot product and still match or surpass Self Attention,If we look at the text generation part of the model what is the effect of Temperature on the probabilities of logits below if an illustrated example of it see next section

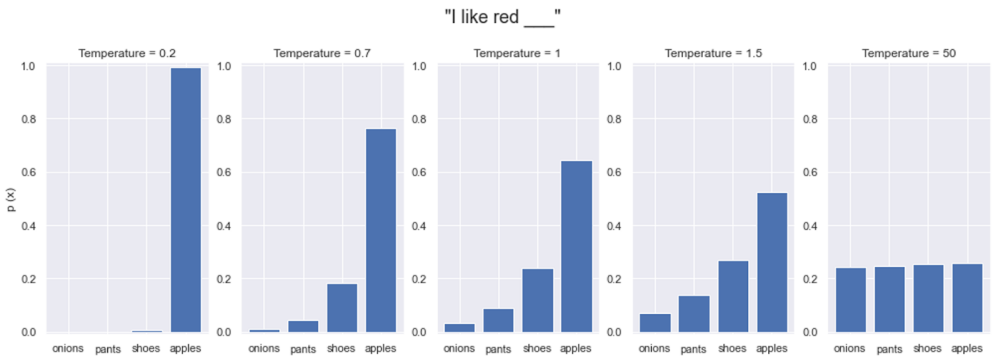


Figure 1: An illustration of how temperature effect the Probabilities.

## 4 Purposed Mechanism

In this section i have proved with initial results that my purposed mechanism outperforms self-attention in every aspect while making the new architecture as simple as possible while remaining efficient to be able for deploying on edge device with limited resources. Below is my purposed modification to the popular architecture used in most Llms

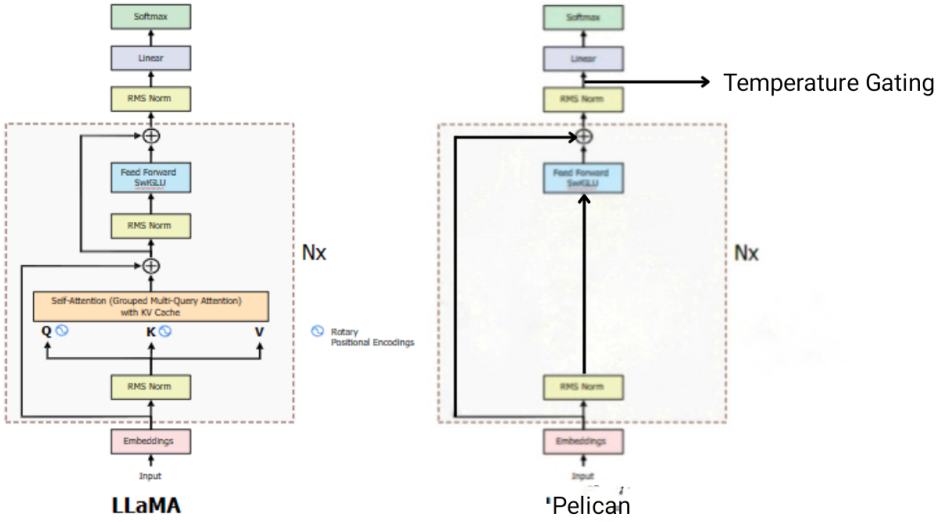


Figure 2: Llama vs Pelican Architecture

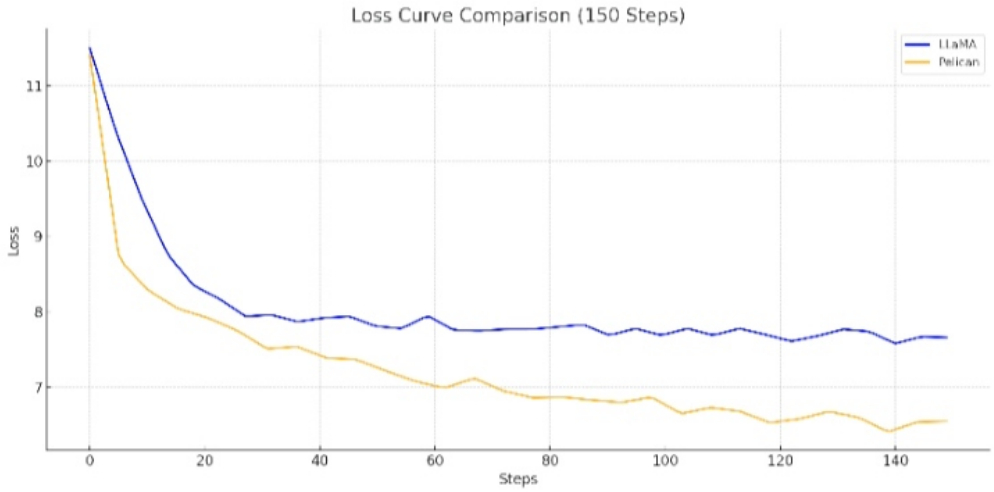


Figure 3: Loss curve of Llama vs pelican on HuggingFaceTB/cosmopedia-100k for 150 steps

Compute complexity of

$$\text{Self-Attention:} \quad \mathcal{O}(n^2 \cdot d)$$
$$\text{Linear Layer:} \quad \mathcal{O}(n \cdot d^2)$$

instead of scaling quadratically it scales linearly with sequence length

As in figure 1: you can note that this mechanism is very similar to self Attention if we lower the temperature values. Generally values between 0.5 and 0.1 performs well beyond the capability of self Attention. This acts as information Gating system
Simply Divide the hidden states from the model Backbone with Temperature value for example 0.2 and pass it to LM Head layer

## 5 Future Directions and Recommendations

To maintain an objective perspective and avoid unintentional hype, the complete loss curve comparison between Pelican and LLaMA has been deliberately omitted. Instead, a link to the GitHub repository containing all experimental results has been provided to ensure transparency and allow for independent verification. Although the current findings are promising, there remains significant potential for refinement and optimization. Future work is encouraged to further explore and enhance the synergy between the core architectural components presented here. Continued research in this direction could lead to more robust and efficient models.

## References

[1] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Anthony Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[3] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.