

Scaling a New Attention Scoring Function: A Feedback on Pretraining a 1B Model Efficiently

Guilhem Fouilhé¹, Niyar R Barman¹, Omar Naim¹, Krish Sharma¹,
Soumadeep Saha¹, Alejandro Estana², and Nicholas Asher^{1,3}

¹IRIT, Toulouse, France

²CALMIP, Toulouse, France

³CNRS, Toulouse, France

Abstract

Understanding the mathematical foundations of modern Large Language Models (LLMs) is essential for designing architectures that are more robust, interpretable, and reliable in real-world settings. However, theoretical analyses often rely on unrealistic assumptions, making large-scale empirical validation necessary.

In previous work, we introduced Scaled Signed Averaging (SSA)[1] as a drop-in replacement for the softmax scoring function in Attention to prevent attention from collapsing to a single token in large/out-of-distribution settings. Simple A/B testing on a 114M parameters model showed promising results but this scale is still far away from the models deployed today.

Through a Mesochallenge/collaboration with CALMIP during the deployment phase of the Kairos supercomputer (64 NVIDIA H200 GPUs, 4.77 Pflop/s, ranked #1 in the GREEN500 November 2025 list), and with Linagora via the OpenLLM France project, we pretrained SSA-Luciole, a 1B-parameter model trained on 3T tokens. This model and training dataset size are closer to current trends. We are currently on the process of transitioning this project to national compute resources for the final stages of the training (continued pretraining with cosine learning rate decay on 1.5T tokens then linear annealing on 0.5T).

This work provides practical feedback on scaling a novel attention scoring function in an academic environment. We discuss the implementation of SSA through custom Triton kernels, optimization challenges encountered during distributed training across 16 nodes, and strategies for monitoring large-scale runs and diagnosing GPU failures. We also present preliminary observations regarding training stability and efficiency compared to standard softmax-based architectures. Results suggests that this architecture replacement could improve any transformer model without any additional computational cost.

Our objective is to provide actionable lessons for academic teams seeking to experiment with novel LLM architectures at scale on modern HPC infrastructures.

References

- [1] Omar Naim, Swarnadeep Bhar, Jérôme Bolte, and Nicholas Asher. SSA: Improving Performance With a Better Scoring Function. *ACL*, 2026.