ReLU's Revival: On the Entropic Overload in Normalization-Free Large Language Models

Nandan Kumar Jha & Brandon Reagen
New York University

Introduction & Motivation

LayerNorm Challenges: Essential for stabilizing LLM training but introduces practical challenges:

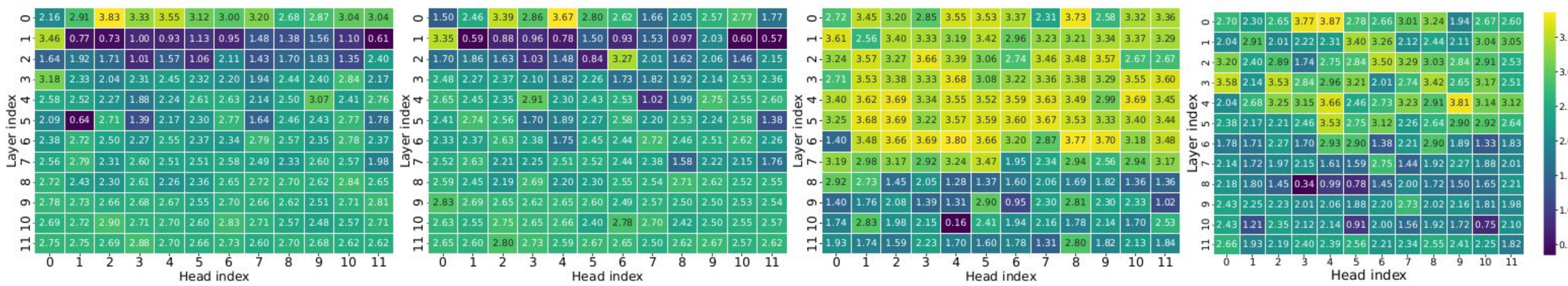
1. Increased complexity in mechanistic interpretability
2. Amplification of outlier features, complicating low-precision training
3. Impaired signal propagation in transformer architectures
4. High latency and communication costs in private inference

Motivation: We explores normalization-free LLM architectures through an *information-theoretic lens*, using *Shannon entropy* to systematically study the impact of FFN activation functions

Key Findings

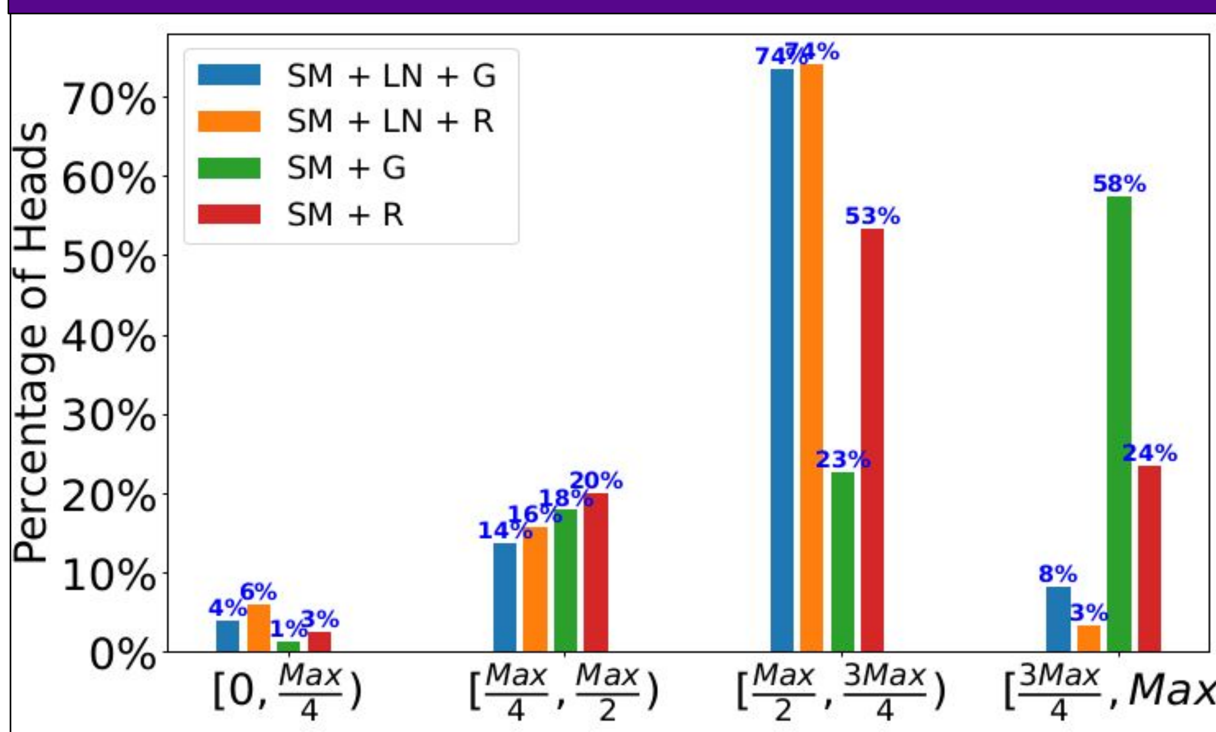
1. ReLU significantly outperforms GELU in LayerNorm-free models (**8.2%** PPL improvement)
2. Early layers in the LayerNorm-Free model with GELU experience *entropic overload*, results in *under-utilization* of MHA's representational capacity
3. LayerNorm-free models naturally converge to ReLU-like behavior with near-zero negative slopes

Entropic Overload in LayerNorm-free model with GELU Activations



(a) SM + LN + G (b) SM + LN + R (c) SM + G (d) SM + R

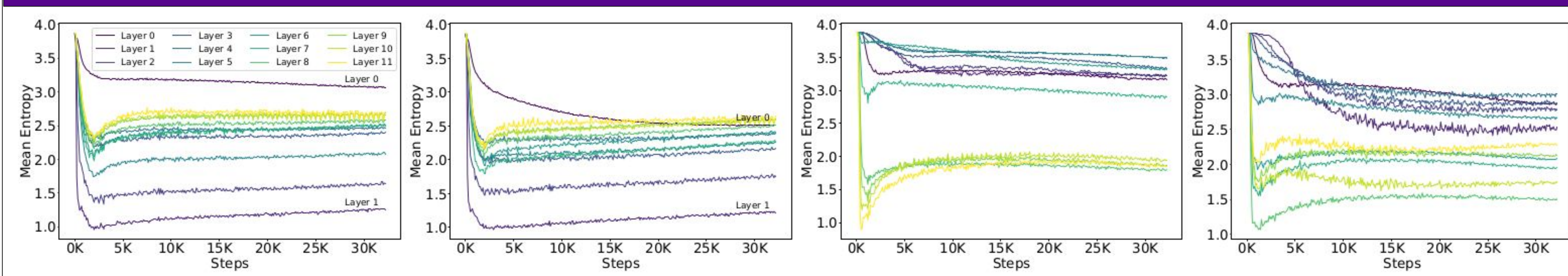
Entropy Distribution



Experimental Results (CodeParrot Dataset, 2.1B Tokens)

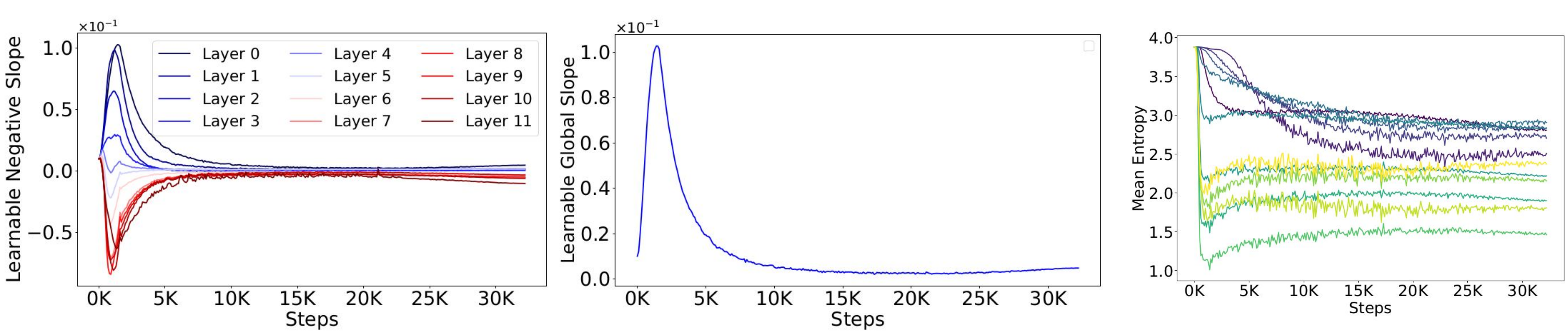
	GPT-2 (T=128)		Pythia-70M (T=128)		Pythia-70M (T=256)	
	Eval PPL	+Δ(%)	Eval PPL	+Δ(%)	Eval PPL	+Δ(%)
SM+LN+G	2.688	0.00	3.512	0.00	3.054	0.00
SM+LN+R	2.757	2.53	3.590	2.22	3.107	1.73
SM+G	3.197	18.92	4.086	16.35	3.570	16.87
SM+R	2.936	9.20	3.736	6.36	3.273	7.17

Layerwise Entropy Dynamics During Pre-training



(a) SM + LN + G (b) SM + LN + R (c) SM + G (d) SM + R

LayerNorm-free Models Naturally Converges to (ReLU-like) Near-Zero Negative Slope



Key Takeaways

1. In LayerNorm-free models, ReLU prevents entropic overload in early layers, enabling **better learning dynamics** and achieving **lower perplexity** compared to GELU.
2. ReLU's geometrical properties, such as **specialization in input space** and **intra-class selectivity**[1], make it naturally effective in the absence of LayerNorm.

1. Alleman et al., Task structure and nonlinearity jointly determine learned representational geometry, ICLR 2024

Paper Code

Contact: nj2049@nyu.edu