

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



3.01 3.24 1.94 2.67 2.6

0 3.26 2.12 2.44 2.11 3.04 3.05

2.01 2.74 3.42 2.65 3.17 2.51

1.38 2.21 2.90 1.89 1.33 1.83

B 1.45 2.00 1.72 1.50 1.65 2.2

00 1.56 1.92 1.72 0.75

6 7 8 9 10 11

2.26 2.64 2.90 2.92 2.6

1.44 1.92 2.27 1.88 2.01

2.73 2.02 2.16 1.81 1.98

2.34 2.55 2.41 2.25 1.8

3.81 3.14 3.12

-3.5

-3.0

-2.5

-2.0

-1.5

-1.0

-0.5

Nandan Kumar Jha & Brandon Reagen **New York University**

Introduction & Motivation

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

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

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

- 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

	Ó	1	ż	3	4	5	6	7	8	9	10	11	
11	2.75	2.75	2.69	2.88	2.70	2.66	2.73	2.60	2.70	2.68	2.62	2.62	11
10	2.69	2.72	2.90	2.71	2.70	2.60	2.83	2.71	2.57	2.48	2.57	2.71	10
6	2.78	2.73	2.66	2.68	2.67	2.55	2.70	2.66	2.62	2.51	2.71	2.81	6
00	2.72	2.43	2.30	2.61	2.26	2.36	2.65	2.72	2.70	2.62	2.84	2.65	00
Lay	2.56	2.79	2:31	2.60	2.51	2.51	2.58	2.49	2.33	2.60	2.57	1.98	Lay
Layer 7 6	2.38	2.72	2.50	2.27	2.55	2.37	2.34	2.79	2.57	2.35	2.78	2.37	yer 6
index 5 4	2.09	0.64	2.71	1.39	2.17	2.30	2.77	1.64	2.46	2.43	2.77	1.78	index 5 4
4 ex	2.58	2.52	2.27	1.88	2.24	2.61	2.63	2.14	2.50	3.07	2.41	2.76	A &
m	3.18	2.33	2.04	2.31	2.45	2.32	2.20	1.94	2.44	2.40	2.84	2.17	m
2	1.64	1.92	1.71	1.01	1.57	1.06	2.11	1.43	1.70	1.83	1.35	2.40	2
-	3.46	0.77	0.73	1.00	0.93	1.13	0.95	1.48	1.38	1.56	1.10	0.61	н
0	2.16	2.91	3.83	3.33	3.55	3.12	3.00	3.20	2.68	2.87	3.04	3.04	0
													_

	0	1	2	3	4 -	-c Head	inde	x	8	9	10	11		0	1	2	3	4	-ead	inde	x	8	9	10	11		U	1	2	3	He	ea
Ч	ó	i	5	ż	À	÷	ė	÷	ò	ó	10	11	Ч	ò	i	5	ż	À	É	ė	÷	8	ó		11	Ч	Å	4	4	5	Å	5
н.	2.65	2.60	2.80	2.73	2.59	2.67	2.65	2.50	2.62	2.67	2.57	2.62	н.	1.93	1.74	1.59	2.23	1.70	1.60	1.78	1.31	2.80	1.82	2.13	1.84	H.	2.66	1 93	2 19	2.40	2.39	21
10	2.63	2.55	2.75	2.65	2.66	2.40	2.78	2.70	2.42	2.50	2.55	2.57	10	1.74	2.83	1.98	2.15	0.16	2.41	1.94	2.16	1.78	2.14	1.70	2.53	10	2.43	1.21	2.35	2.12	2.14	0.9
б	2.83	2.69	2.65	2.62	2.65	2.60	2.49	2.57	2.50	2.50	2.53	2.54	6	1.40	1.76	2.08	1.39	1.31	2.90	0.95	2.30	2.81	2.30	2.33	1.02	6	2.43	2.25	2.23	2.01	2.06	1.8
ω.	2.59	2.45	2.19	2.69	2.20	2.30	2.55	2.54	2.71	2.62	2.52	2.55	00	2.92	2.73	1.45	2.05	1.28	1.37	1.60	2.06	1.69	1.82	1.36	1.36	00	2.18	1.80	1.45	0.34	0.99	0.
-	2.52	2.63	2.21	2.25	2.51	2.52	2.44	2.38	1.58	2.22	2.15	1.76	La	3.19	2.98	3.17	2.92	3.24	3.47	1.95	2.34	2.94	2.56	2.94	3.17	Lay	2.14	1.72	1.97	2.15	1.61	1.
			2.63			1	10 - D						>	2	S	3.66					2 3			÷		-	and the second second	Contraction of the second		and the second second	2.93	
		10000		10.000	1.000			100 C	10000		1999 - 1999 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 -	1.38			10000000	3.69					Transfer of the	1000	10000	Sector Sector	1.000	ind	2.38	2.17	2.21	2.46	3.53	2.
	1		2.35			2	1				1		· •			3.69										-	1.1.1.1.1.1.1	1.		- C - C - C - C - C - C - C - C - C - C	3.66	1.5
21.112.0		10-100 J	2.37	Sector Sector	1.000				10000		Section.	10000000			and the second second	3.38				All the second second	and the second	1000	1000			m	3.58	2.14	3.53	2.84	2.96	3.2
	1 1		1.63	S 1	-		1			11 12						3.27	_				1			1	51 L	2	3.20	2.40	2.89	1.74	2.75	2.8
-	3.35	0.59	0.88	0.96	0.78	1.50	0.93	1.53	0.97	2.03	0.60	0.57	Ч	3.61	2.56	3.40	3.33	3.19	3.42	2.96	3.23	3.21	3.34	3.37	3.29	ч	2.04	2.91	2.01	2.22	2.31	3/
0	1.50	2.46	3.39	2.86	3.67	2.80	2.62	1.66	2.05	2.57	2.77	1.77	0	2.72	3.45	3.20	2.85	3.55	3.53	3.37	2.31	3.73	2.58	3.32	3.36	0	2.70	2.30	2.65	3.77	3.87	2.
			-						_		_			_			_						_				_					

Head index

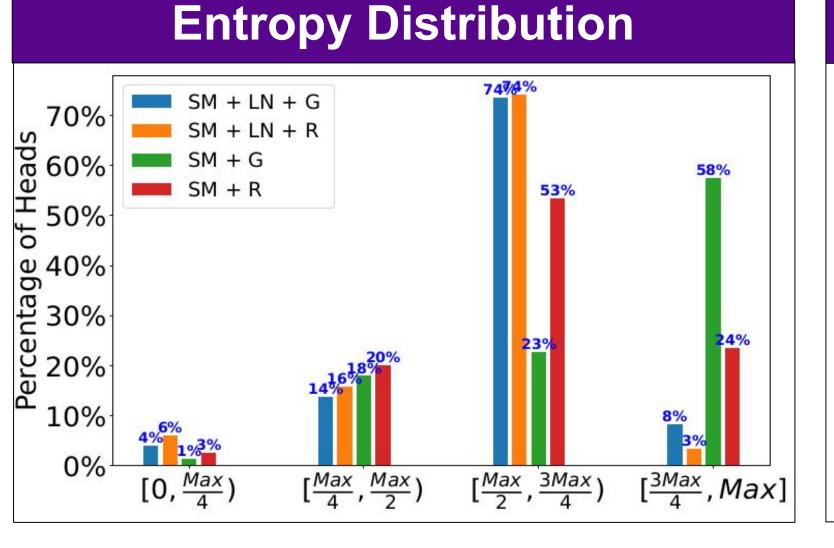
(a) SM + LN + G

(b) SM + LN + R

(c) SM + G

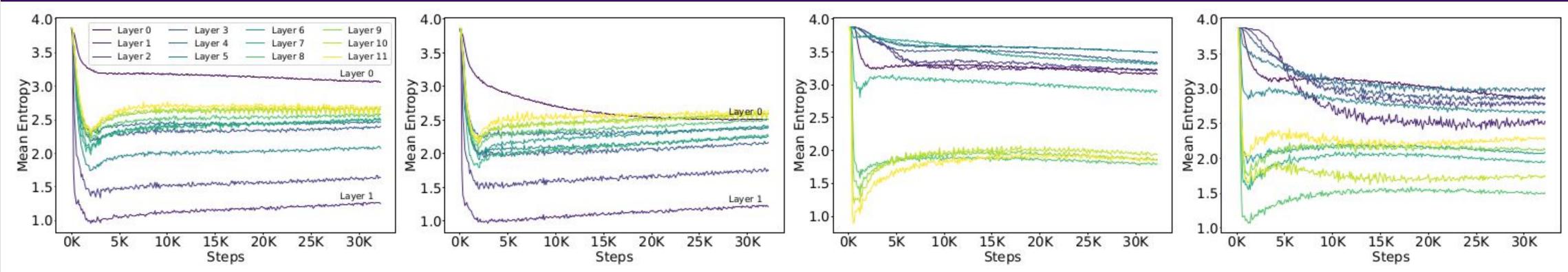
ead index

(d) SM + R



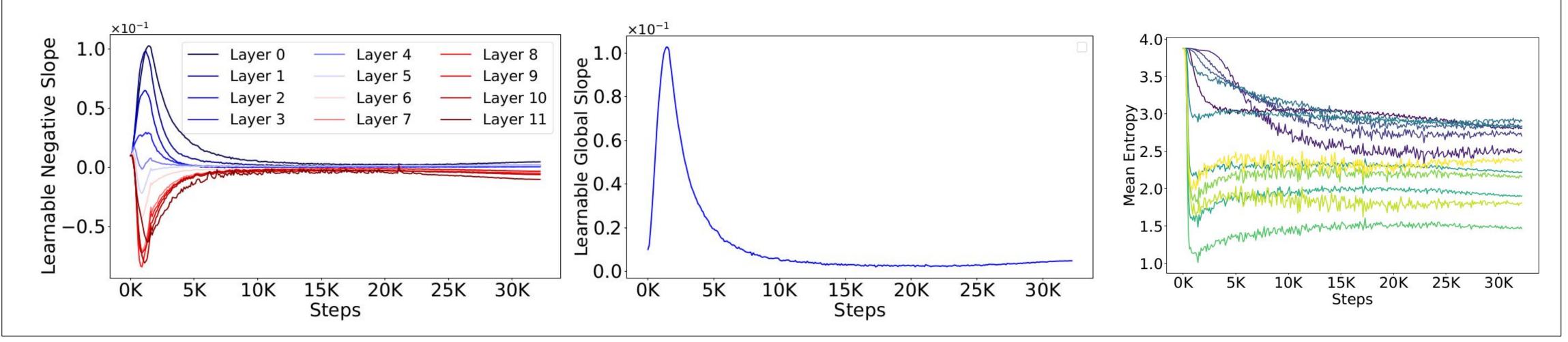
Experin	nental Re	sults (C	odeParro	t Datase	t, 2.1B Tol	kens)
	GPT-2 (7	7=128)	Pythia-70M	I (T=128)	Pythia-70M	I (T=256)
	Eval PPL	$+\Delta(\%)$	Eval PPL	$+\Delta(\%)$	Eval PPL	$+\Delta(\%)$
SM+LN+G SM+LN+R	2.688 2.757	0.00 2.53	3.512 3.590	0.00 2.22	3.054 3.107	0.00 1.73
SM+G SM+R	3.197 2.936	18.92 9.20	4.086 3.736	16.35 6.36	3.570 3.273	16.87 7.17

Layerwise Entropy Dynamics During Pre-training



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

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



Key Takeaways

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

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

