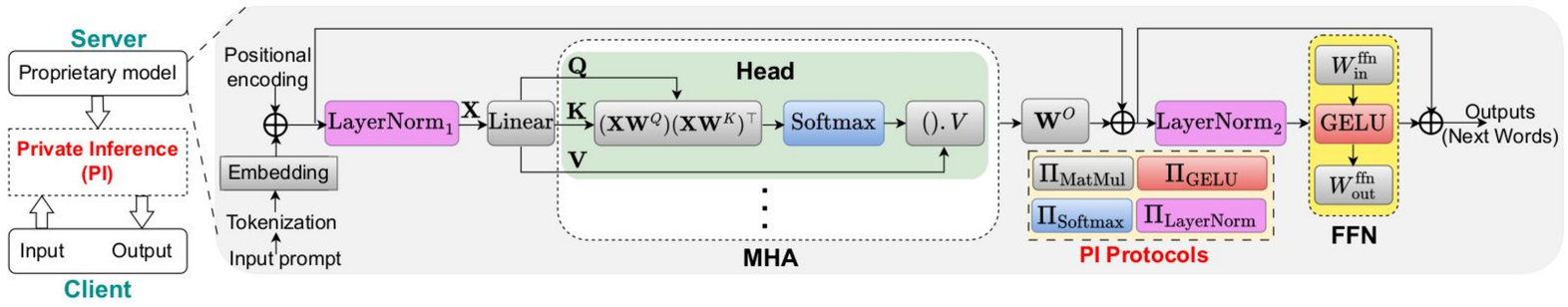
Entropy-Guided Attention for Private LLMs

Nandan Kumar Jha & Brandon Reagen
New York University

The 39th Annual AAAI Conference On Artificial Intelligence

Private Inference on Large Language Models: An Introduction

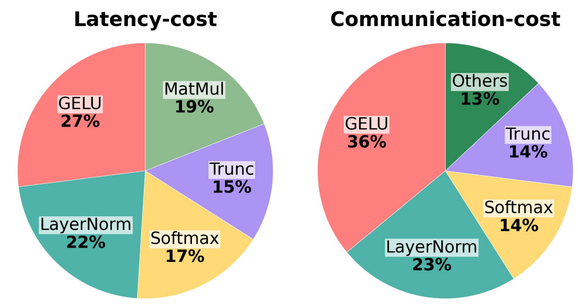


Computations performed directly on encrypted data, without seeing it's content

Privacy of users' sensitive data is preserved while the server's model remains protected

Motivation and Challenges

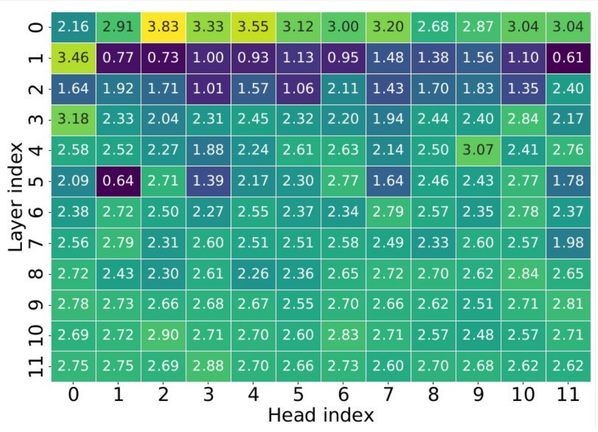
High Latency & Communication Overheads: **8.2** minutes and **25.3** GB to generate a single token on GPT-2 (125M, T=128)

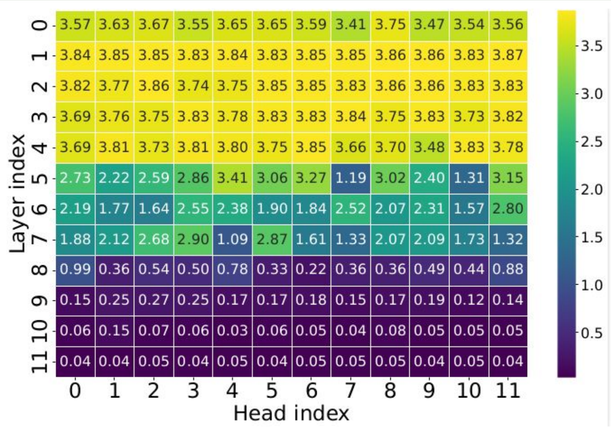


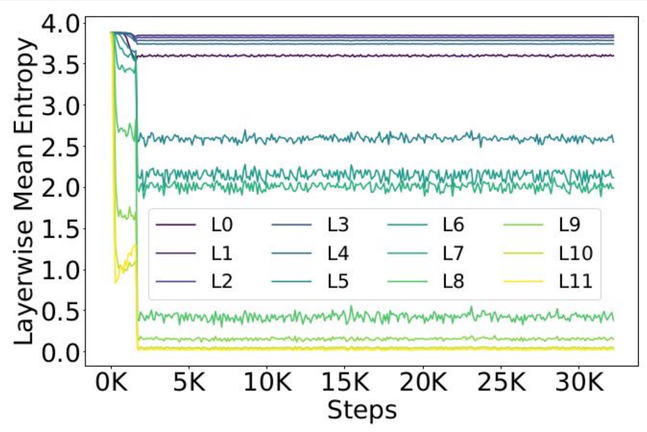
Nonlinearity Bottleneck [1]

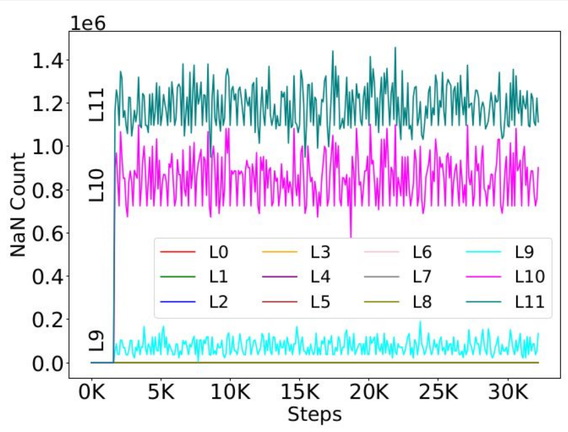
1. Hou et al., CipherGPT: Secure Two-Party GPT Inference

Key Findings: The Absence of Nonlinearities in LLMs Leads to Entropic Overload & Entropy Collapse









Balanced entropies with nonlinearities

Without LayerNorm & GELU: **Entropic overload** (reducing heads' diversity) & **entropy collapse** (training instability)

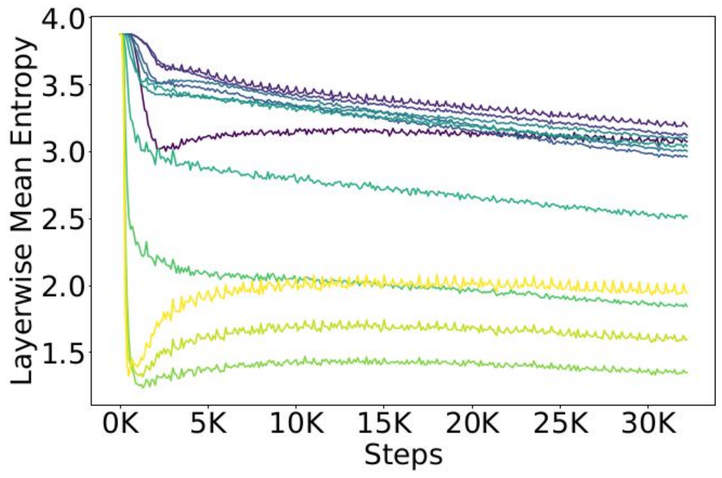
Inference-Efficient Solutions to Prevent Entropy Collapse

Solution 1: Weight Normalization in FFN

$$\text{FFN}^{\text{SM}}_{\text{WNorm}}(\mathbf{X}) = \left(\mathbf{X}\frac{\mathbf{V}^{\text{ffn}}_{\text{in}}}{\|\mathbf{V}_{\text{in}}\|_2}g_{\text{in}}\right)\frac{\mathbf{V}^{\text{ffn}}_{\text{out}}}{\|\mathbf{V}_{\text{out}}\|_2}g_{\text{out}}$$

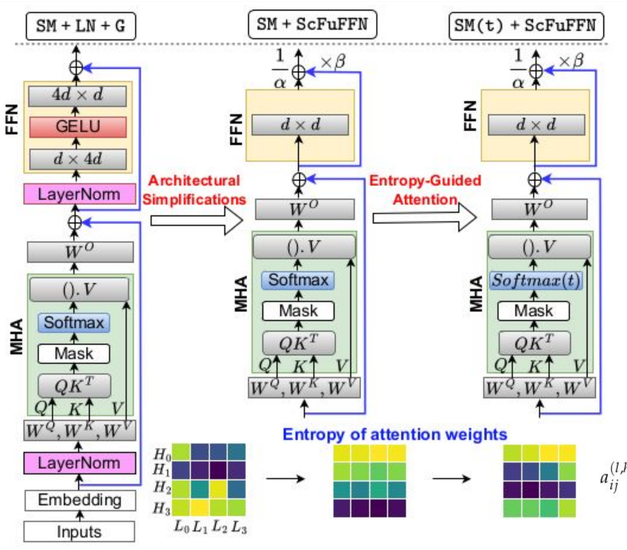
Solution 2: Spectral Normalization in FFN

$$\text{FFN}^{\text{SM}}_{\text{SNorm}}(\mathbf{X}) = \left(\mathbf{X}\frac{\mathbf{W}^{\text{ffn}}_{\text{in}}}{\sigma(\mathbf{W}^{\text{ffn}}_{\text{in}})}\right)\frac{\mathbf{W}^{\text{ffn}}_{\text{out}}}{\sigma(\mathbf{W}^{\text{ffn}}_{\text{out}})}$$



Solution 3: Learnable-Scaling in FFN

Entropy-Guided Private LLM Design



Scaled FFN

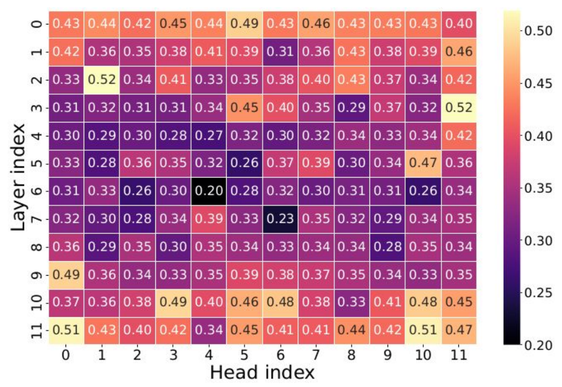
$$\mathbf{X}_{\text{out}} = \beta\hat{\mathbf{X}}_{\text{SA}} + \frac{1}{\alpha}(\text{FFN}^{\text{SM}}(\mathbf{X}_{\text{SA}}))$$

Parameterized Attention

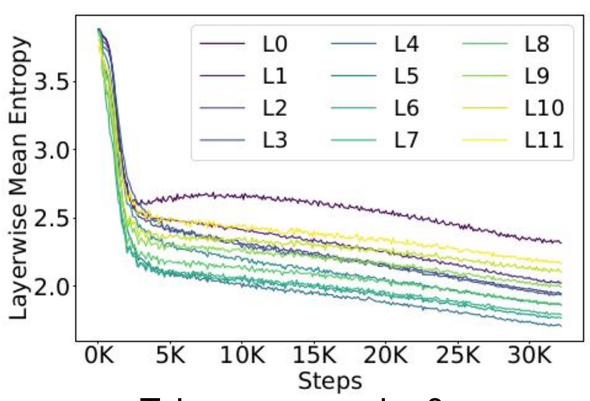
$$\mathbf{A}^{(l,h)}(t) = \left[a^{(l,h)}_{ij}(t)\right]_{T\times T}$$

$$a^{(l,h)}_{ij}(t) = \frac{\exp\left(\frac{1}{t_i\sqrt{d_k}}(\mathbf{X}_i\mathbf{W}^Q)(\mathbf{X}_j\mathbf{W}^K)^\top\right)}{\sum_{k=1}^{T}\exp\left(\frac{1}{t_i\sqrt{d_k}}(\mathbf{X}_i\mathbf{W}^Q)(\mathbf{X}_k\mathbf{W}^K)^\top\right)}$$

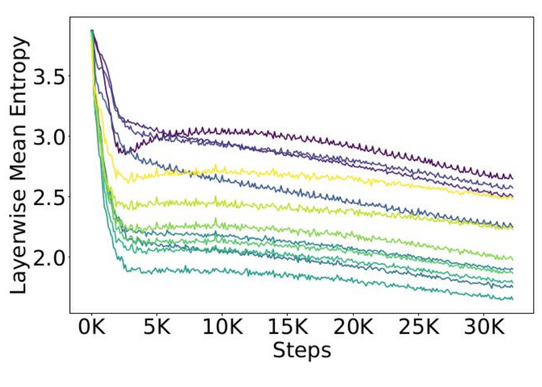
Key Innovation in Entropy Regularization Scheme: Learnable Threshold and Tolerance Margin



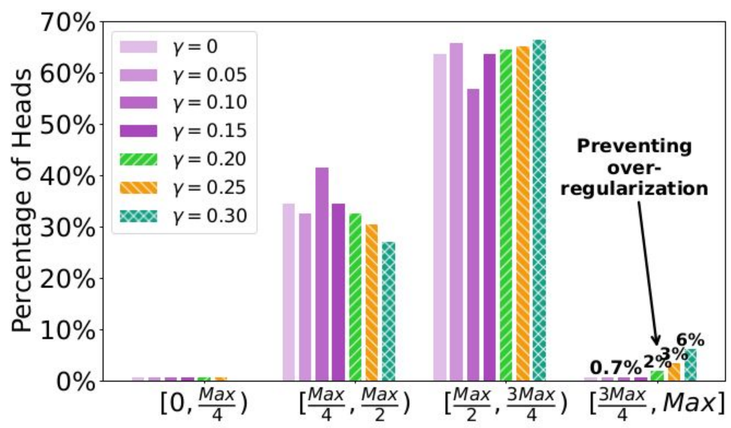
Head-wise learnable thresholds



Tolerance-margin: 0



Tolerance-margin: 0.30E_{max}



Experimental Results: GPT-2 (L=12, H=12, d=768)

	Network Arch.	PPL	#Nonlinear Ops	#FLOPs FFN	#FLOPs Attn.	Comm. (GB)	Lat. (min.)	Savings Comm.	Savings Lat.
Baseline	SM + LN + G	2.69	SM:144 × $\mathbb{R}^{128\times128}$ LN:24 × $\mathbb{R}^{128\times768}$ G:12 × $\mathbb{R}^{128\times3072}$	14.5B	7.7B	25.32	8.21	1×	1×
Baseline	SM + LN + R	2.76	SM:144 × $\mathbb{R}^{128\times128}$ LN:24 × $\mathbb{R}^{128\times768}$ R:12 × $\mathbb{R}^{128\times3072}$	14.5B	7.7B	9.44	6.06	2.68×	1.35×
	SM + ScFuFFN	3.48	SM:144 × $\mathbb{R}^{128\times128}$	1.8B	7.7B	6.43	4.76	3.94×	1.72×
	EReg(SM(t) + ScFuFFN)	3.21	SM:144 × $\mathbb{R}^{128\times128}$	1.8B	7.7B	6.43	4.76	3.94×	1.72×

CodeParrot Dataset (2.1B Tokens, T=128)

	Network Arch.	Eval PPL 1.2B	Eval PPL 2.4B	Eval PPL 4.8B	#Nonlinear Ops	#FLOPs FFN	#FLOPs Attn.	Comm. (GB)	Lat. (min.)
Baseline	SM + LN + G	25.71	23.32	21.29	SM:144 × $\mathbb{R}^{512\times512}$ LN:24 × $\mathbb{R}^{512\times768}$ G:12 × $\mathbb{R}^{512\times3072}$	58.0B	36.2B	145.24	30.74
Baseline	SM + LN + R	26.06	23.55	21.58	SM:144 × $\mathbb{R}^{512\times512}$ LN:24 × $\mathbb{R}^{512\times768}$ R:12 × $\mathbb{R}^{512\times3072}$	58.0B	36.2B	81.71	23.54
	SM + ScFuFFN	33.77	30.82	28.59	SM:144 × $\mathbb{R}^{512\times512}$	7.3B	36.2B	69.68	19.44
	EReg(SM(t) + ScFuFFN)	31.54	28.70	26.55	SM:144 × $\mathbb{R}^{512\times512}$	7.3B	36.2B	69.68	19.44

Languini Book Dataset (1.2B to 4.8B Tokens, T=512)

Conclusion and Key Takeaways

While nonlinearities are essential in LLMs, **strategically** applied entropy regularization and FFN normalization can prevent entropy collapse and entropic overload in (Softmax-only) private LLMs.

Paper

Code

Contact:
nj2049@nyu.edu