OPT2025: 17th Annual Workshop on Optimization for Machine Learning

Regularizing the Entropy Landscape of Self-Attention: Towards a Soft Inductive Bias in LLMs

Nandan Kumar Jha NJ2049@NYU.EDU
New York University

Brandon Reagen BJR5@NYU.EDU
New York University

Abstract

Self-attention often looks under-utilized: many heads can be pruned with little loss. We revisit this inefficiency through the entropy landscape of multi-head attention and ask whether a *soft inductive bias* can steer optimization toward more useful regimes. To this end, we employ a head-wise entropy regularizer with learnable per-head strengths and optional softmax temperatures that penalize only excess entropy. On decoder-only language models (e.g., GPT-2), this simple training-time prior improves perplexity by up to **20%** without inference overhead. Internally, it reshapes the entropy profile: early layers shift toward lower entropy, the high-entropy tail disappears, and head-role heterogeneity increases, reducing overlap among heads. Our findings suggest that standard training induces a high-entropy attractor in early self-attention layers and convergence toward homogeneous attention values in deeper layers. A soft entropic bias gently redirects this path, transforming redundancy into functional specialization while keeping inference cost unchanged.

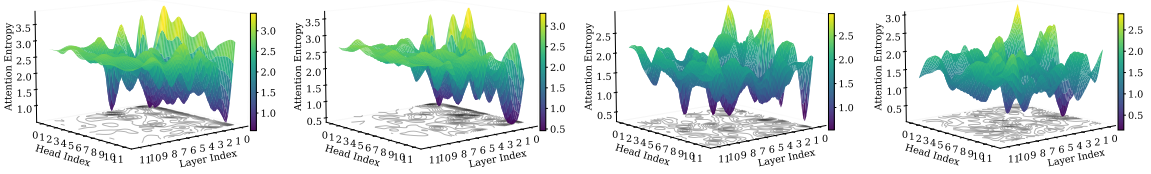
1. Introduction

Multi-head attention (MHA) is the computational backbone of Transformers, yet it often appears over-parameterized: large fractions of heads can be pruned with little impact [9–12]. This inefficiency raises a deeper question: *why do models under-recruit their attention capacity, and can training-time priors encourage healthier utilization?*

We revisit this problem through the **entropy landscape** of self-attention. Under standard training recipes (softmax attention with cross-entropy loss, optimized with AdamW), early layers consistently exhibit high-entropy attention: distributions spread broadly across tokens, heads overlap, and functional specialization is diluted. Our analysis uncovers three recurring suboptimal patterns: 1) elevated mean entropy, 2) persistent high-entropy tails, and 3) reduced heterogeneity across heads, indicating that optimization is biased toward diffuse regimes.

To address this, we turn to the notion of **soft inductive bias** [1, 3, 4, 13], which emphasizes priors that guides optimization rather than restrict the hypothesis class. Unlike hard biases such as architectural constraints or structured attention, soft biases are permissive: they guide training toward more useful solutions while preserving expressivity.

We instantiate this principle in MHA by employing a head-wise entropy regularizer, where each head is equipped with learnable strengths and softmax temperatures, together with a global tolerance margin that penalizes extreme entropies while allowing natural diversity [7, 8]. Importantly, this acts as a soft inductive bias: it does not constrain the hypothesis class but nudges optimization toward healthier entropy regimes. Across decoder-only LLMs, this simple training-time prior



(*a*) GPT2(GELU) (*b*) GPT2(ReLU) (*c*) GPT2(GELU)+EReg (*d*) GPT2(ReLU)+EReg

Figure 1: Attention-entropy landscapes across layers and heads for GPT-2 (125M) variants. Each surface shows the mean attention entropy per head, averaged over tokens at context length (T=128). Higher entropy indicates diffuse attention; lower entropy indicates selective attention. *Applying entropy regularization shifts the head-wise entropy profile toward greater heterogeneity and a lower global mean entropy*.

yields consistent gains. Perplexity improves by up to **20%** (Table 1), while internal organization also changes: early layers shift to lower entropy (Figure 1), high-entropy tails vanish (Figure 3), and head-role heterogeneity increases (Figure 2). These effects reduce redundancy and recover functional diversity of heads.

Our contributions are threefold: (i) an **entropy-landscape analysis** revealing an optimization-induced high-entropy attractor in self-attention, (ii) a **head-wise soft entropic bias** that is inference-neutral, and (iii) empirical evidence that this bias improves perplexity and functional specialization. Together, these results highlight entropy as a lens for understanding attention utilization and demonstrate that subtle regularization can redirect optimization, turning redundancy into specialization.

2. Experimental Setup

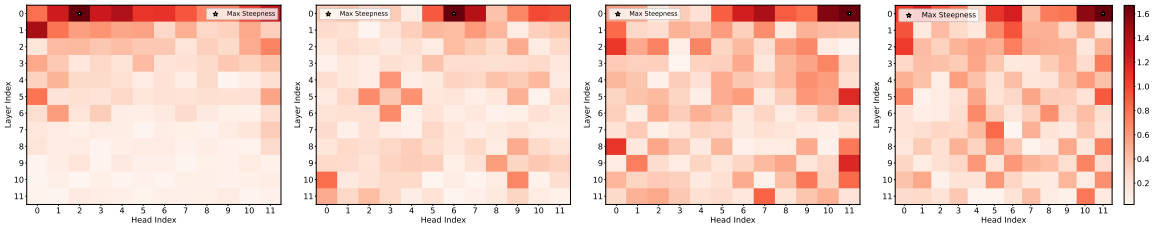
To evaluate the role of soft inductive bias in shaping attention dynamics, we train GPT-2 models with 12 and 18 layers on the CodeParrot dataset [2], a widely used benchmark for large language models [5, 6]. CodeParrot is sourced from approximately 20 million Python files on GitHub and contains 8 GB of code with 16.7 million examples. Each example consists of 128 tokens, yielding a total of 2.1 billion training tokens. We employ a Byte-Pair Encoding tokenizer with a 50K vocabulary and train with context lengths of 128 and 256. The structural redundancy of source code makes CodeParrot a natural testbed: it stresses model capacity and provides a clear lens to study whether a soft entropic bias can reduce redundancy, promote functional diversity, and improve overall perplexity.

3. Regularizing the Attention Entropy Landscape

Notations. We denote the number of layers as L, number of heads as H, model dimensionality as d, head dimension as d_k (where $d_k = \frac{d}{H}$), and context length as T.

Let $\mathbf{A}^{(h,l)} \in \mathbb{R}^{T\times T}$ be the attention matrix of h-th head in l-th layer, and each element in the attention matrix, $a_{ij}^{(l,h)}$, are attention weights for the i-th query and j-th key, which are non-negative and sum to one for a query:

$$\mathbf{A}^{(l,h)} = \left[a_{ij}^{(l,h)}\right]_{T\times T}, \quad \text{where} \quad a_{ij}^{(l,h)} \geq 0 \quad \text{and} \quad \sum_{j=1}^{T} a_{ij}^{(l,h)} = 1 \tag{1}$$



(*a*) Baseline (*b*) EReg($T_{\text{margin}} = 0$) (*c*) EReg($T_{\text{margin}} = 0.10\times$) (*d*) EReg($T_{\text{margin}} = 0.20\times$)

Figure 2: Landscape steepness of attention entropy. Heatmap of the gradient magnitude $|\nabla E|$ over the layer-head grid, where E is mean token-wise attention entropy. Bright regions mark sharp transitions—either across heads (specialization boundaries) or across layers (mass shift). **(a) Baseline** shows relatively uniform, low-gradient regions in deeper layers (layers 6-11), indicating homogeneous attention patterns. **(b-d) Entropy regularization** with increasing target margins (T_{margin} = 0, 0.20E_{max}, 0.20E_{max}) progressively increases heterogeneity, creating sharper attention head specialization boundaries.

This square matrix is generated by applying the softmax operation over the key length for each query position as follows

$$\mathbf{A}^{(h,l)}(\mathbf{X}) = \text{Softmax}\Big(\frac{1}{\sqrt{d_k}}(\mathbf{X}\mathbf{W}^Q)(\mathbf{X}\mathbf{W}^K)^\top\Big), \text{ where } \quad \text{Softmax}(\mathbf{X}_i) = \frac{\exp(x_i)}{\sum_{j=1}^{T}\exp(x_j)} \tag{2}$$

Following [14], we compute the mean of entropy values across all query positions to obtain a single entropy value for each head. The entropy $\mathbf{E}^{(l,h)}$ for the h-th head in the l-th layer of an attention matrix is given by:

$$\mathbf{E}^{(l,h)} = -\frac{1}{T}\sum_{i=1}^{T}\sum_{j=1}^{T} a_{ij}^{(l,h)} \log\left(a_{ij}^{(l,h)} + \epsilon\right), \quad \text{where} \quad a_{ij}^{(l,h)} = \frac{\exp\left(\frac{1}{\sqrt{d_k}}(\mathbf{X}_i\mathbf{W}^Q)(\mathbf{X}_j\mathbf{W}^K)^\top\right)}{\sum_{k=1}^{T}\exp\left(\frac{1}{\sqrt{d_k}}(\mathbf{X}_i\mathbf{W}^Q)(\mathbf{X}_k\mathbf{W}^K)^\top\right)} \tag{3}$$

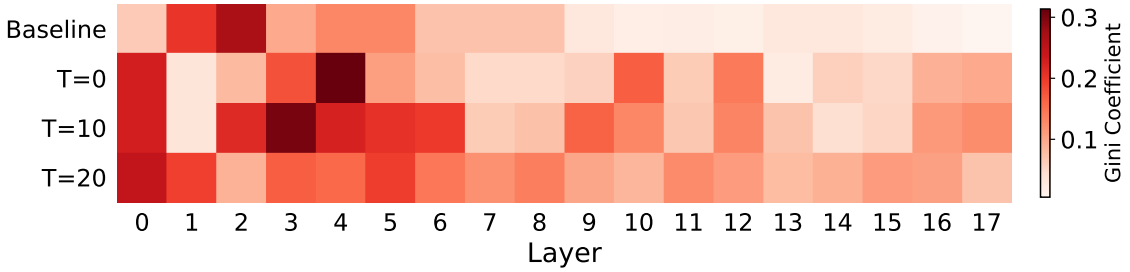
where ϵ is a small constant added for numerical stability to prevent taking the log of zero.

4. Experimental Results

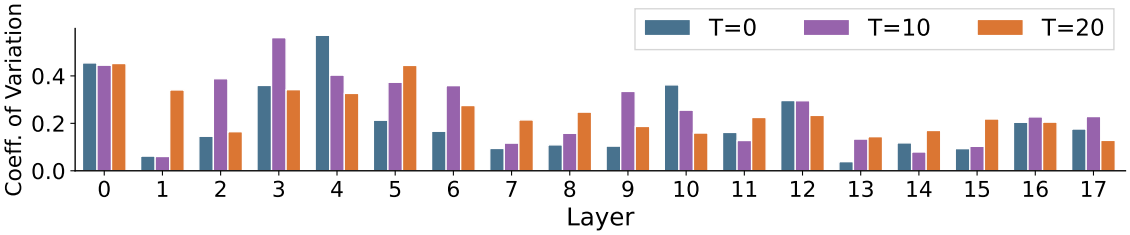
4.1. Manifold Topology Analysis of Attention Entropy Landscapes

To characterize the geometric structure of attention entropy manifolds across the transformer's representational hierarchy, we analyze the gradient field $|\nabla E| = \sqrt{(\frac{\partial E}{\partial \text{layer}})^2 + (\frac{\partial E}{\partial \text{head}})^2}$, where E denotes the median token-wise attention entropy. This differential geometry approach quantifies the local curvature and discontinuities in the attention entropy surface, revealing the topological organization of attention behavioral modes. High gradient magnitude regions correspond to sharp phase transitions in attention focusing regimes, indicative of representational bottlenecks or functional specialization boundaries, while low-gradient regions signify smooth manifold regions with consistent attention mechanistic properties.

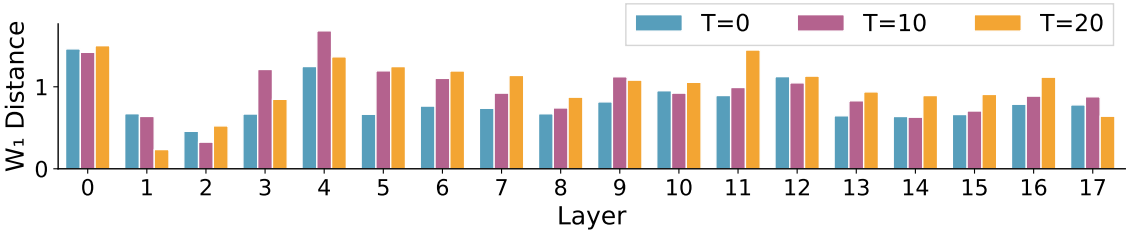
Our gradient field analysis exposes a critical failure mode in standard transformer architectures: attention head collapse in deeper layers. Baseline models exhibit pathologically low gradient magnitudes in layers 6-11 (Figure 2), indicating that the attention entropy manifold becomes increasingly flat and degenerate with depth. This manifold collapse suggests that attention heads undergo



(*a*) Coefficient of variation heatmap showing attention diversity across layers and conditions



(*b*) Gini coefficients measuring attention inequality for each regularization strength



(*c*) Wasserstein distances quantifying distributional mass shift from baseline attention patterns

Figure 3: Layerwise analysis of functional diversity and heterogeneity of attention heads under different regularization strengths in GPT-2 models with 18 layers. Entropy regularization progressively improves the attention diversity and shifts the mean of the distributions from baseline across the layers.

representational convergence, losing their capacity for diverse computational specialization, a phenomenon analogous to feature map redundancy in convolutional networks. The gradient magnitude tensor visually manifests this degeneracy as sparse, low-energy regions in the deeper representational subspace, revealing the extent of attention head homogenization that undermines the model's expressivity.

Entropy regularization fundamentally reshapes the attention manifold topology through controlled diversification pressure, inducing sharp gradient boundaries that prevent representational collapse (Figure 2). As the regularization strength increases, we observe progressive manifold refinement with heightened local curvature, particularly along intra-layer head boundaries. This topological restructuring demonstrates that entropy regularization operates not merely as a distributional constraint, but as a geometric prior that enforces attention head orthogonality in the entropy space. The resulting high-gradient regions indicate successful attention head disentanglement, where adjacent mechanisms develop orthogonal computational roles. This gradient-based analysis provides

Table 1: Performance comparison across different tolerance margin settings in entropy regularization. PPL denotes perplexity, Ent. Range shows entropy range bounds, and W1-Dist. represents Wasserstein-1 distance.

		Baseline		**Tmargin=0**			**Tmargin=10E$_{max}$**			**Tmargin=20E$_{max}$**		
		PPL	Ent. Range	PPL	Ent. Range	W1-Dist.	PPL	Ent. Range	W1-Dist.	PPL	Ent. Range	W1-Dist.
L=12, T=128	GELU	2.688	[0.62, 3.84]	2.189	[0.01, 2.57]	0.789	2.222	[0.21, 2.64]	0.856	2.169	[0.31, 2.97]	0.878
	ReLU	2.757	[0.58, 3.68]	2.395	[0.07, 2.62]	0.656	2.381	[0.12, 2.63]	0.750	2.367	[0.27, 3.03]	0.842
L=18, T=128	GELU	2.655	[0.07, 3.85]	2.168	[0.00, 2.67]	0.813	2.185	[0.07, 2.77]	0.958	2.127	[0.03, 2.97]	1.007
	ReLU	2.625	[0.26, 3.87]	2.264	[0.01, 2.64]	0.726	2.229	[0.10, 2.73]	0.863	2.198	[0.03, 2.83]	0.882
L=12, T=256	ReLU	2.638	[0.30, 4.55]	2.288	[0.02, 3.26]	0.677	2.224	[0.06, 3.33]	0.726	2.207	[0.16, 3.70]	0.812

direct evidence that entropy regularization prevents the notorious attention head redundancy problem by maintaining sufficient manifold complexity throughout the network depth.

4.2. Attention Diversity Under Entropy Regularization

Entropy regularization achieves substantial attention diversification. Across all model configurations, regularization systematically expands attention entropy ranges (e.g., from [0.62, 3.84] to [0.31, 2.97] for GELU L=12) while maintaining competitive perplexity (Table 1). Wasserstein-1 distances of 0.656-1.007 confirm fundamental distributional shifts in attention patterns rather than minor adjustments (Figure 3). This performance-diversity trade-off proves consistent across architectural variations, activation functions (GELU/ReLU), model depths (L=12/18), and context lengths (T=128/256). The robustness suggests entropy regularization operates through architecture-agnostic attention head specialization mechanisms, providing principled expressivity enhancement without sacrificing language modeling capability.

5. Conclusion

We revisited the inefficiency of multi-head attention through the lens of its entropy landscape and showed that standard training drifts toward diffuse, high-entropy regimes that undermine specialization. To resolve this, we introduced a soft entropic bias nudging optimization without altering inference. Across GPT-2 models, this approach yields up to 20% perplexity improvement, while reshaping internal organization by suppressing high-entropy tails and enhancing head-role heterogeneity. These findings highlight entropy as a powerful diagnostic and design tool for attention, and suggest that subtle regularization can recover under-utilized LLM capacity.

References

[1] Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *International conference on machine learning (ICML)*, 2021.

[2] Hugging Face. Codeparrot. https://huggingface.co/learn/nlp-course/chapter7/6.

[3] Marc Finzi, Gregory Benton, and Andrew G Wilson. Residual pathway priors for soft equivariance constraints. In *Advances in Neural Information Processing Systems*, 2021.

[4] Micah Goldblum, Marc Anton Finzi, Keefer Rowan, and Andrew Gordon Wilson. Position: the no free lunch theorem, kolmogorov complexity, and the role of inductive biases in machine learning. In *Forty-first International Conference on Machine Learning (ICMl)*, 2024.

[5] Bobby He and Thomas Hofmann. Simplifying transformer blocks. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[6] Bobby He, Lorenzo Noci, Daniele Paliotta, Imanol Schlag, and Thomas Hofmann. Understanding and minimising outlier features in neural network training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[7] Nandan Kumar Jha and Brandon Reagen. AERO: Softmax-only llms for efficient private inference. *arXiv preprint arXiv:2410.13060*, 2024.

[8] Nandan Kumar Jha and Brandon Reagen. Entropy-guided attention for private llms. *arXiv preprint arXiv:2501.03489*, 2025.

[9] Jae-young Jo and Sung-Hyon Myaeng. Roles and utilization of attention heads in transformer-based neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[10] Weicheng Ma, Kai Zhang, Renze Lou, Lili Wang, and Soroush Vosoughi. Contributions of transformer attention heads in multi- and cross-lingual tasks. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.

[11] Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in neural information processing systems*, 2019.

[12] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

[13] Andrew Gordon Wilson. Position: Deep learning is not so mysterious or different. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.

[14] Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua M Susskind. Stabilizing transformer training by preventing attention entropy collapse. In *International Conference on Machine Learning (ICML)*, 2023.