

Rethinking Attention: Entropy-Guided Iterative Refinement for Language Models

Subham Divakar^{1,*} and Rojalina Priyadarshini^{2,*}

¹ Department of Electronics and Communication Engineering, C.V. Raman Global University, Bhubaneswar 752054, India

² Department of Computer Science and Engineering, C.V. Raman Global University, Bhubaneswar 752054, India

* Correspondence: shubham.divakar@gmail.com (S.D.); priyadarshini.rojalina@gmail.com (R.P.)

How To Cite: Divakar, S.; Priyadarshini, R. Rethinking Attention: Entropy-Guided Iterative Refinement for Language Models. *Transactions on Artificial Intelligence* 2026, 2(1), 214–230. https://doi.org/10.53941/tai.2026.100014

Received: 26 July 2026

Revised: 29 August 2026

Accepted: 10 September 2026

Published: 29 September 2026

Abstract: Standard Multi-Head Attention (MHA) computes contextual representations in a single forward pass, providing no mechanism to detect or correct diffuse or suboptimal attention distributions. Existing efforts to improve attention have primarily targeted computational efficiency or sequence length, leaving attention quality itself largely unaddressed. We propose IMHA (Entropy-Guided Iterative Multi-Head Attention), a within-layer attention formulation that progressively refines token representations over T iterations using shared projection weights. At each step, the Shannon entropy of each token’s per-head attention row—normalised against the position-dependent maximum to remove a structural artefact of causal masking—serves as an intrinsic quality signal: tokens with peaked (low-entropy) attention are amplified while those with diffuse (high-entropy) attention are suppressed. A gated residual update incorporating a prefix-causal global context vector ensures stable convergence. The formulation is strictly causal, verified at the gradient level across all ablation configurations. We prove (Proposition 1) that under mild assumptions the token-specific content of an attention output decays monotonically with attention diffuseness, providing principled motivation for entropy weighting. We further show (Proposition 2) that if iterative refinement sharpens the attention distribution, discriminative content is non-decreasing across iterations. We evaluate across three model scales (Small 14 M, Medium 45 M, Large 87 M parameters) on WikiText-2, Penn Treebank, and TinyStories, with downstream transfer to SST-2 sentiment classification and AG News topic classification. At medium scale, IMHA reduces WikiText-2 perplexity from 47.30 ± 0.18 to 44.85 ± 0.16 over the MHA baseline (paired t -test, $p = 3.2 \times 10^{-6}$, $d = 11.4$, after Holm–Bonferroni correction over 18 comparisons), outperforms parameter-matched capacity baselines by 1.40 points, and surpasses attention-quality methods including α -entmax, RealFormer, Universal Transformer (FLOPs-matched), and entropy-regularised MHA. Critically, IMHA outperforms entropy-regularised MHA by 1.18 points, demonstrating that the *internal*, within-layer delivery of the entropy signal—rather than entropy as an external loss term—is what drives the improvement. The relative gain is stable across scales (4.8%, 4.9%, 4.6% at Small, Medium, and Large respectively), and the pretrained representations transfer to classification: IMHA achieves $93.2 \pm 0.3\%$ on SST-2 vs. $91.8 \pm 0.4\%$ for MHA, and $90.1 \pm 0.4\%$ vs. $88.7 \pm 0.5\%$ on AG News. Total overhead is a uniform $\sim 8\%$ across parameters, FLOPs, training time, and inference latency.

Keywords: transformer attention; iterative refinement; entropy-based attention; language modeling; information theory; representation quality; transfer learning

1. Introduction

1.1. The Fundamental Limitation of Single-Pass Attention

The transformer architecture [1] and its Multi-Head Attention (MHA) mechanism have become the foundation of modern natural language processing [2,3]. The core operation of MHA—computing pairwise token interactions



through softmax-normalised dot products—is simultaneously the source of its expressive power and its central limitation: attention distributions are determined in a single forward step from the current representations, and once computed, they are never revisited or evaluated for quality.

This matters because attention distributions in practice span a wide range of quality. Some tokens attend sharply to a small set of contextually relevant positions, producing representations rich in token-specific information. Others spread weight nearly uniformly, producing representations that approach the mean of all available value vectors and carry little discriminative content. Standard MHA treats both situations identically: after softmax normalisation, all tokens contribute to the output weighted solely by pairwise similarity, with no mechanism to detect which attention patterns are informative and which are noise.

The consequences are measurable. Voita et al. [4] demonstrated that in trained transformer models, the majority of attention heads learn redundant or structurally uninformative patterns, with only a small subset capturing linguistically meaningful functions. Jain and Wallace [5] showed that attention weights may not reliably reflect token importance in downstream predictions. These findings collectively suggest that the quality of attention distributions, not just their computation, is a lever that has been left on the table.

1.2. Why Existing Approaches Do Not Solve This

A large and growing body of research addresses limitations of standard attention, but along three directions that leave attention quality largely untouched:

Efficiency methods reduce the computational cost of attention through approximation, hashing, or kernel reformulation [6–11]. These make attention faster or more memory-efficient while computing essentially the same operation representationally.

Context-extension methods enable attention over longer sequences [12–15]. These extend where attention can reach but not the quality of the distributions formed over that context.

Full-block iterative methods such as Universal Transformer [16] and Deep Equilibrium Models [17] iterate over the entire transformer block (attention + feed-forward + layer normalisation) with shared weights. While these demonstrate the value of iterative computation, they treat all sub-layers identically and do not leverage information-theoretic properties of the attention distribution to guide refinement.

Crucially, none of these approaches uses the uncertainty structure of the attention distribution itself—as measured by Shannon entropy [18]—as an intrinsic quality signal. This is the gap IMHA addresses.

1.3. The IMHA Approach

We propose IMHA (Entropy-Guided Iterative Multi-Head Attention), which departs from standard MHA in three specific ways:

- (1) Iterative refinement with shared projections. Rather than computing attention once, IMHA re-derives queries, keys, and values from progressively updated representations over T iterations using a single shared set of projection matrices. This allows coarse initial attention estimates to be refined as representations sharpen, without the parameter overhead of separate weights per iteration.
- (2) Entropy-guided token weighting. At each iteration, the Shannon entropy of each token’s per-head attention distribution is computed and normalised against the position-dependent maximum (removing a structural artefact of causal masking). This normalised entropy serves as a per-token confidence score: tokens with peaked attention (low normalised entropy) are amplified; tokens with diffuse attention (high normalised entropy) are suppressed.
- (3) Gated residual update with prefix-causal global context. Token representations are updated through a sigmoid gate that incorporates both the entropy-weighted local output and a prefix-causal global context vector, ensuring stable convergence while maintaining the autoregressive property.

1.4. The Key Distinguishing Experiment

A natural question is whether the benefits of IMHA come from the information-theoretic signal (entropy) or merely from providing additional computational budget and a regularisation effect. We answer this directly by including entropy-regularised MHA (MHA with $-\lambda \cdot H_{\text{mean}}$ added to the training loss) as a baseline. This variant uses the same entropy signal as IMHA but delivers it externally as a loss term rather than internally as a within-layer reweighting mechanism.

IMHA outperforms entropy-regularised MHA by 1.18 points on WikiText-2 at medium scale. This gap isolates the contribution of the delivery mechanism: it is not sufficient to push the model to reduce attention entropy globally

via a loss term. The benefit comes from using entropy within the layer, at the level of individual token-head pairs, to selectively amplify informative representations in real time during the forward pass.

1.5. Contributions

This paper makes the following specific contributions:

- (1) We propose IMHA, a strictly prefix-causal attention mechanism that introduces per-head entropy-guided token weighting and within-layer iterative refinement, with causality verified at the gradient level across all ablation configurations.
- (2) We prove Proposition 1 (Section 4): under mild assumptions, the token-specific content of an attention output decays monotonically with attention diffuseness, providing formal motivation for entropy-based reweighting.
- (3) We prove Proposition 2 (Section 4): if iterative refinement sharpens the attention distribution, discriminative content is non-decreasing across iterations.
- (4) We demonstrate that IMHA's gains are attributable to the internal delivery of the entropy signal, not merely to the entropy signal itself, through comparison with entropy-regularised MHA.
- (5) We evaluate IMHA across three model scales (14–87 M parameters), three language modeling benchmarks, two downstream classification tasks, and a comprehensive ablation grid, with full statistical reporting (paired tests, exact p -values, CIs, Cohen's d , Holm–Bonferroni correction).

1.6. Paper Organisation

Section 2 surveys related work. Section 3 describes the IMHA formulation in detail. Section 4 provides the theoretical analysis. Section 5 presents the component ablation. Section 6 describes datasets, scales, and baselines. Section 7 reports experimental results. Section 8 analyses why IMHA works. Section 9 concludes with limitations and future directions.

2. Related Work

2.1. Efficient Attention Mechanisms

Reducing the quadratic cost of attention has been a major research focus. Linear attention [6] approximates softmax with kernel functions for $\mathcal{O}(N)$ complexity. Performer [7] uses random feature maps to approximate the softmax kernel. Linformer [8] projects keys and values to lower dimension. Reformer [9] uses locality-sensitive hashing. Nyströmformer [10] applies low-rank matrix approximation. FlashAttention [11] improves memory efficiency through IO-aware tiling while computing exact softmax attention. Routing Transformers [19] use content-based sparse attention. A comprehensive survey is given by Tay et al. [20].

All of these methods operate at the implementation or approximation level; they do not address the quality of the attention distributions that result. FlashAttention in particular produces numerically identical outputs to standard MHA and is included in our efficiency comparison only.

2.2. Context Extension

Transformer-XL [12] enables attention beyond fixed context windows through segment-level recurrence. Longformer [13] and BigBird [14] use sparse attention patterns to scale to longer documents. Adaptive attention span [15] learns per-head context windows. These methods determine what attention can see, not the quality of the distributions formed over what it sees.

2.3. Attention Distribution Refinement

The closest line of related work directly modifies the attention distribution. Correia et al. [21] replace softmax with α -entmax, producing sparse distributions that assign exact zero weight to irrelevant tokens. This is a single-pass modification to the normalisation function. IMHA differs in retaining softmax and using entropy computed from the distribution as a multi-step reweighting signal.

He et al. [22] proposed RealFormer, which adds residual connections from one layer's attention logits to the next layer's scores. This is a cross-layer refinement; IMHA's refinement is within-layer, guided by an explicit quality signal rather than raw logit residuals.

Entropy-based regularisation of attention has been explored in several forms, most commonly as an auxiliary loss that penalises uniform or excessively peaked attention distributions [23,24]. The key distinction from IMHA is the delivery mechanism: a loss term enforces an average entropy constraint across training, but does not

selectively amplify high-quality tokens during the forward pass. We test this distinction experimentally via the entropy-regularised MHA baseline.

In computer vision, Divakar and Priyadarshini [25] demonstrated that recursive attention refinement improves feature selection in CNNs. The present work extends this principle to the transformer setting with information-theoretic guidance.

2.4. Iterative and Recurrent Transformers

Universal Transformer [16] iterates the entire transformer block (attention + FFN + LN) with shared weights, using Adaptive Computation Time (ACT) [26] to learn per-position halting. IMHA differs in three ways: (i) it iterates only the attention sub-layer; (ii) entropy replaces a learned halting criterion; (iii) it uses a fixed small number of iterations ($T = 2$) rather than adaptive depth. For fair comparison we include Universal Transformer with FLOPs matched to IMHA (Section 6).

Deep Equilibrium Models (DEQ) [17] find fixed points of weight-tied layers via implicit differentiation. IMHA uses a fixed T with standard backpropagation and targets only the attention sub-layer.

2.5. Positioning

Table 1 places IMHA among the most closely related methods. IMHA occupies a unique position: within-layer scope, entropy-guided signal, and iterative refinement of the attention sub-layer only.

Table 1. Positioning of IMHA. The combination of within-layer scope, entropy as an intrinsic quality signal, and attention-sub-layer iteration is unique.

Method	Unit Iterated	Refinement Signal	Scope
Universal Transformer	Full block	Learned halting	Cross-layer
DEQ	Full block	Fixed-point residual	Cross-layer
RealFormer	None	Prior layer logits	Cross-layer
α -entmax	None	Normalisation fn.	Within-layer
Entropy-reg. MHA	None	Loss term	Training
IMHA (ours)	Attn. sub-layer	Attn. entropy	Within-layer

2.6. Information Theory and Representations

Shannon entropy [18] provides the conceptual foundation. The information bottleneck framework [27] argues that optimal representations compress irrelevant variation while retaining task-relevant signal—precisely the property IMHA promotes in attention outputs. Prior work on attention entropy has documented that heads with lower entropy tend to be more interpretable and task-relevant [4], providing empirical support for entropy as an attention quality signal.

3. Methodology

3.1. Notation

$X \in \mathbb{R}^{N \times d}$: input sequence of N tokens with dimension d . H attention heads with per-head dimension $d_k = d/H$. Causal mask $M_{ij} = 0$ for $j \leq i$ and $-\infty$ for $j > i$. σ denotes sigmoid, $[\cdot]$ feature-dimension concatenation, $\odot$ element-wise multiplication. All logarithms are natural.

3.2. Standard Multi-Head Attention and Its Limitation

Standard MHA computes, for each head h :

$$Q^{(h)} = XW_Q^{(h)}, K^{(h)} = XW_K^{(h)}, V^{(h)} = XW_V^{(h)} \quad (1)$$

$$A^{(h)} = \text{softmax}\left(\frac{Q^{(h)}K^{(h)\top}}{\sqrt{d_k}} + M\right), z^{(h)} = A^{(h)}V^{(h)} \quad (2)$$

and returns $[z^{(1)}; \dots; z^{(H)}]W_O$. The single-pass design means attention quality is fixed by the initial representations X . A diffuse attention row—spreading weight uniformly over context—yields a representation $z_i \approx \mu$ (the mean value vector), which carries no token-specific information. Standard attention has no way to detect or correct this situation.

3.3. IMHA Formulation

3.3.1. Iterative Refinement with Shared Projections

Let $h^{(0)} = X$. At each iteration $t \in \{0, \dots, T-1\}$:

$$Q^{(t,h)} = h^{(t)} W_Q^{(h)}, K^{(t,h)} = h^{(t)} W_K^{(h)}, V^{(t,h)} = h^{(t)} W_V^{(h)} \quad (3)$$

$$A^{(t,h)} = \text{softmax}\left(\frac{Q^{(t,h)} K^{(t,h)\top}}{\sqrt{d_k}} + M\right) \quad (4)$$

$$z^{(t,h)} = A^{(t,h)} V^{(t,h)} \quad (5)$$

The projections $\{W_Q^{(h)}, W_K^{(h)}, W_V^{(h)}\}$ are shared across all T iterations. IMHA therefore introduces zero additional projection parameters. Sharing also acts as implicit regularisation: the model must learn projections that produce useful representations at every iteration stage.

3.3.2. Per-Head Entropy and Position Normalisation

For token i , iteration t , head h :

$$H_i^{(t,h)} = - \sum_{j=0}^i A_{ij}^{(t,h)} \log A_{ij}^{(t,h)} \quad (6)$$

using $0 \log 0 = 0$. Under causal masking, token i attends over exactly $i + 1$ positions, so $H_i \in [0, \log(i + 1)]$ by construction. Raw entropy is thus position-dependent by necessity, not by choice. We normalise:

$$\hat{H}_i^{(t,h)} = \frac{H_i^{(t,h)}}{\log(i + 1)} \in [0, 1] \quad (7)$$

This measures diffuseness relative to what is attainable at position i , making the signal comparable across positions. At $i = 0$, we set $\hat{H}_0 = 0$.

3.3.3. Prefix-Causal Entropy Weighting

We convert normalised entropies to per-token weights using a prefix mean:

$$w_i^{(t,h)} = \frac{\exp(-\hat{H}_i^{(t,h)}/\tau)}{\frac{1}{i+1} \sum_{k=0}^i \exp(-\hat{H}_k^{(t,h)}/\tau)} \quad (8)$$

where $\tau > 0$ is a learnable temperature per layer (initialised at 1). Using the prefix mean rather than the prefix sum as normaliser keeps $\mathbb{E}[w_i] \approx 1$ at every position, avoiding the $\mathcal{O}(1/i)$ decay that a prefix softmax would produce. The denominator ranges over $k \leq i$ only, so w_i depends on no future token.

Four properties hold by construction: (i) causality, (ii) mean-one scale invariance at every position, (iii) sensitivity controlled by τ , (iv) $\mathcal{O}(N)$ computation via a single cumulative sum.

The weighted per-head output is:

$$\tilde{z}_i^{(t,h)} = w_i^{(t,h)} \cdot z_i^{(t,h)} \quad (9)$$

3.3.4. Multi-Head Concatenation

Entropy is computed per head. After entropy weighting, per-head outputs are concatenated through the shared output projection:

$$z_i^{(t)} = [\tilde{z}_i^{(t,1)}; \dots; \tilde{z}_i^{(t,H)}] W_O \quad (10)$$

$W_O \in \mathbb{R}^{d \times d}$ is shared across iterations.

3.3.5. Prefix-Causal Global Context

The below Equation (11) represents prefix-causal global context computed via cumulative sum and is discussed in depth below.

$$\bar{g}_i^{(t)} = \frac{1}{i+1} \sum_{k=0}^i z_k^{(t)} \quad (11)$$

computed via cumulative sum in $\mathcal{O}(N)$. Each token receives a summary of its own prefix. This provides sequence-level context without violating autoregressive constraints.

3.3.6. Gated Residual Update

The gated residual mechanism is the one which controls the magnitude of each iterative refinement step. For each token, a gate determines how much of the candidate refinement should be added into the current hidden representation. The gate and candidate update are computed as follows:

$$\alpha_i^{(t)} = \sigma(W_\alpha[z_i^{(t)}; \bar{g}_i^{(t)}]) \quad (12)$$

$$f_i^{(t)} = \text{ReLU}(W_f[z_i^{(t)}; \bar{g}_i^{(t)}]) \quad (13)$$

$$h_i^{(t+1)} = h_i^{(t)} + \alpha_i^{(t)} \odot f_i^{(t)} \quad (14)$$

The gate $\alpha_i^{(t)} \in (0, 1)^d$ prevents overshooting. If a token's attention is already peaked, the gate learns to apply a small update. The design is analogous to the GRU update gate [28].

$W_\alpha, W_f \in \mathbb{R}^{2d \times d}$ are shared across iterations but separate from the attention projections. They are the only additional parameters relative to standard MHA.

3.3.7. Output Contract and Causality

IMHA returns $h^{(T)} - h^{(0)}$, so the block computes $x \leftarrow x + \text{IMHA}(\text{LN}(x))$, identical to standard MHA. IMHA at $T = 1$ with entropy, global context, and gating disabled is numerically identical to standard MHA (max diff 1.19×10^{-7}).

Causality is verified at three levels: (1) structural—Equations (8) and (11) sum over $k \leq i$ only; (2) perturbation—changing input at position j leaves every output at $i < j$ bitwise unchanged across all 96 ablation configurations; (3) gradient—automatic differentiation confirms $\partial h_i^{(T)} / \partial x_j = 0$ exactly for $j > i$.

3.3.8. Full Algorithm

The complete forward pass is given in Algorithm 1. For-loops are presented for clarity; all operations are batched GPU matrix computations.

Algorithm 1 IMHA Forward Pass

Require: $X \in \mathbb{R}^{N \times d}$; shared $W_Q^{(h)}, W_K^{(h)}, W_V^{(h)}, W_O$; W_α, W_f ; τ ; T ; mask M

Ensure: Residual update $h^{(T)} - h^{(0)}$

```

1:  $h^{(0)} \leftarrow X$ 
2: for  $t = 0$  to  $T - 1$  do
3:   for head  $h = 1$  to  $H$  (parallel) do
4:     Compute  $A^{(t,h)}$  via Equations (3) and (4)
5:      $H_i^{(t,h)} \leftarrow -\sum_j A_{ij}^{(t,h)} \log A_{ij}^{(t,h)}$ 
6:      $\hat{H}_i^{(t,h)} \leftarrow H_i^{(t,h)} / \log(i+1)$ 
7:      $w_i^{(t,h)} \leftarrow \exp(-\hat{H}_i^{(t,h)} / \tau) / \text{prefix-mean}$ 
8:      $\tilde{z}_i^{(t,h)} \leftarrow w_i^{(t,h)} \cdot z_i^{(t,h)}$ 
9:   end for
10:   $z^{(t)} \leftarrow [\tilde{z}^{(t,1)}; \dots; \tilde{z}^{(t,H)}] W_O$ 
11:   $\bar{g}_i^{(t)} \leftarrow \frac{1}{i+1} \sum_{k \leq i} z_k^{(t)}$ 
12:   $h_i^{(t+1)} \leftarrow h_i^{(t)} + \sigma(W_\alpha[z_i^{(t)}; \bar{g}_i^{(t)}]) \odot \text{ReLU}(W_f[z_i^{(t)}; \bar{g}_i^{(t)}])$ 
13: end for
14: return  $h^{(T)} - h^{(0)}$ 

```

3.3.9. Overhead Analysis

Additional parameters: $W_\alpha, W_f \in \mathbb{R}^{2d \times d}$ per layer. At $d = 512$, $L = 6$: +3.1 M parameters (+7.0%). Additional computation: one extra attention matmul ($\mathcal{O}(N^2 d_k H)$) and two gating matmuls ($\mathcal{O}(N d^2)$) per layer. Measured overhead: uniform $\sim 8\%$ across parameters, FLOPs, training time, and latency.

4. Theoretical Analysis

We provide two results that formally motivate the IMHA design. Both are motivating analyses under stated assumptions, not optimality guarantees.

4.1. Assumptions

Fix token i . Write $z_i = \sum_{j \leq i} A_{ij} v_j$, and model value vectors as:

$$v_j = \mu + \varepsilon_j, \quad \mathbb{E}[\varepsilon_j] = \mathbf{0}, \quad \text{Cov}(\varepsilon_j) = \sigma_v^2 I \quad (15)$$

(A1). *Bounded variance*: $\sigma_v^2 < \infty$.

(A2). *Pairwise uncorrelated deviations*: $\mathbb{E}[\varepsilon_j^\top \varepsilon_k] = 0$ for $j \neq k$.

(A3). *Empirical H – H_2 monotonicity*: Shannon entropy H and Rényi-2 entropy H_2 are monotonically related on the model's attention distributions. Tested directly in Section 8.3.

(A2) idealises independence; we discuss its relaxation in Section 4.4.

4.2. Proposition 1: Diffuseness Attenuates Signal

The attention distribution determines how much token specific information is retained in the output representation. When the attention weights are concentrated on a small number of tokens, the weighted aggregation preserves larger deviations from the prefix mean. In contrast, when the attention distribution becomes diffuse, contributions from individual tokens are averaged more strongly, causing the output to approach the context-independent mean. The following proposition formalizes this relationship.

Proposition 1 (Entropy and discriminative content). *Under (A1)–(A2), squared deviation of the attention output from the prefix mean satisfies:*

$$\mathbb{E}\|z_i - \mu\|^2 = \sigma_v^2 \sum_{j \leq i} A_{ij}^2 = \sigma_v^2 \exp(-H_2(A_i)) \quad (16)$$

where $H_2(A_i) = -\log \sum_j A_{ij}^2$ is the Rényi-2 entropy. $\mathbb{E}\|z_i - \mu\|^2$ is strictly decreasing in H_2 : as attention becomes more diffuse, z_i collapses towards the context-independent mean.

Proof. Since $\sum_{j \leq i} A_{ij} = 1$, we have $z_i - \mu = \sum_{j \leq i} A_{ij} \varepsilon_j$. Taking expectations:

$$\mathbb{E}\|z_i - \mu\|^2 = \sum_{j,k} A_{ij} A_{ik} \mathbb{E}[\varepsilon_j^\top \varepsilon_k] \stackrel{(A2)}{=} \sigma_v^2 \sum_{j \leq i} A_{ij}^2 = \sigma_v^2 \exp(-H_2(A_i))$$

Monotonicity follows from $\exp(-\cdot)$ being strictly decreasing. $\square$

Corollary 1 (Extreme cases). (i) *Uniform attention*: $\mathbb{E}\|z_i - \mu\|^2 = \sigma_v^2/(i+1) \rightarrow 0$. No token-specific content. (ii) *One-hot attention*: $\mathbb{E}\|z_i - \mu\|^2 = \sigma_v^2$. Full token-specific deviation retained.

Proposition 1 shows that $\exp(-\hat{H}_i)$, used as a token weight in Equation (8), is a monotone proxy for the discriminative content of that token's representation (under A3). High-content tokens are amplified; low-content tokens are suppressed.

Remark 1 (Correction of a prior formulation). *An earlier version of this analysis characterised $\sum_j A_{ij}^2$ as the variance of the attention estimate and concluded high entropy implies high variance. That is inverted: $\sum_j A_{ij}^2 = \exp(-H_2)$ is largest when attention is peaked. The relevant object is the magnitude of token-specific content retained, which has the opposite dependence on entropy. We thank the reviewers for identifying this.*

4.3. Proposition 2: Iterative Refinement Preserves Content

While Proposition 1 characterizes how attention diffuseness affects the amount of token-specific content retained by a single attention operation, the proposed iterative architecture repeatedly applies the same refinement process. It is therefore important to establish whether this recursion necessarily causes the representation to lose the content preserved by the initial aggregation. Under a residual refinement formulation, the iterative updates can preserve the existing representation while progressively incorporating additional contextual information.

Proposition 2 (Refinement and content). *Under (A1)–(A3), if the mean normalised entropy decreases across iterations ($\mathbb{E}[\hat{H}^{(t+1)}] \leq \mathbb{E}[\hat{H}^{(t)}]$), then the mean expected discriminative content is non-decreasing across iterations.*

Proof. By Proposition 1, $\mathbb{E}\|z_i - \mu\|^2 = \sigma_v^2 \exp(-H_2(A_i))$ is monotone decreasing in H_2 . Under (A3), H_2 decreases monotonically with H . If $\hat{H}^{(t+1)} \leq \hat{H}^{(t)}$ in expectation, then $H_2^{(t+1)} \leq H_2^{(t)}$ in expectation by (A3), so $\exp(-H_2^{(t+1)}) \geq \exp(-H_2^{(t)})$, meaning content is non-decreasing. $\square$

Proposition 2 establishes that the iterative refinement is not circular: if the mechanism does in fact sharpen attention distributions (an empirical question verified in Section 8.3), then content preservation follows from the theory. The antecedent is an empirical condition; we measure the entropy trajectory directly.

4.4. Scope and Limitations of the Analysis

Assumption (A2) does not hold exactly in trained models: semantically related tokens have correlated value vectors. Under positive correlation, the cross terms in the variance expansion are positive for all non-degenerate attention patterns, but the qualitative ordering (more diffuse $\Rightarrow$ closer to μ) is preserved when diagonal terms dominate—likely when the number of attended positions is moderate. We treat the analysis as motivating rather than dispositive, and identify a tighter treatment via mutual information or PAC-Bayes bounds as future work.

5. Ablation Study

We conduct a controlled ablation study to quantify the contribution of each component of IMHA and to distinguish the effects of iterative refinement, entropy-based weighting, global context, and adaptive gating.

The component grid considers both single-component variants, which measure the contribution of each mechanism in isolation, and leave-one-out variants, which measure the degradation incurred when a component is removed from the complete IMHA architecture. We further evaluate alternative design choices for entropy normalization and temperature scaling.

5.1. Component Grid

All ablations run at Small config (14 M parameters) on WikiText-2 with five shared seeds are shown in Table 2. This isolates the mechanism from scale effects.

Table 2. Component ablation, WikiText-2 (mean $\pm$ std, 5 seeds). p_{adj} : Holm–Bonferroni vs. full IMHA. $p_{\text{adj}} > 0.05$: not significantly different from full model.

Variant	Test PPL	p_{adj}
Standard MHA	50.12 $\pm$ 0.21	1.4×10^{-4}
Only one component (all others off):		
Iteration only ($T = 2$, no entropy/global/gate)	49.35 $\pm$ 0.19	3.8×10^{-4}
Entropy only ($T = 1$)	49.05 $\pm$ 0.18	1.2×10^{-3}
Global context only ($T = 1$)	49.68 $\pm$ 0.22	7.5×10^{-4}
Gating only ($T = 1$)	49.50 $\pm$ 0.20	4.1×10^{-4}
Full IMHA minus one component:		
- Entropy	48.92 $\pm$ 0.20	6.3×10^{-3}
- Global context	48.35 $\pm$ 0.18	2.8×10^{-2}
- Gating	48.48 $\pm$ 0.17	1.8×10^{-2}
- Iteration ($T = 1$)	48.60 $\pm$ 0.19	9.1×10^{-3}
Design choices:		
Unnormalised entropy (raw H)	48.42 $\pm$ 0.21	2.2×10^{-2}
Head-averaged entropy	48.10 $\pm$ 0.18	0.31
Fixed $\tau = 1$	48.08 $\pm$ 0.19	0.42
Full IMHA	47.85 $\pm$ 0.19	—

Key Findings

Entropy weighting is the single most impactful component (1.07 points solo vs. 0.77 for iteration), consistent with Proposition 1. All components are complementary: the full model outperforms every leave-one-out variant. Position normalisation matters ($p_{\text{adj}} = 0.022$), confirming the structural artefact described in Section 3.3.2 has practical impact. Per-head entropy and learnable τ do not differ significantly from their simpler alternatives, but we retain them as defaults since they add no cost.

5.2. Effect of Refinement Iterations

Table 3 shows the effect of T (refinement iterations) and its results over the various steps.

Table 3. Effect of T (mean $\pm$ std, 5 seeds). $T = 2$ is optimal across all datasets; degradation beyond $T = 2$ rules out simple compute scaling.

T	WikiText-2	PTB	TinyStories
1	48.60 $\pm$ 0.19	53.40 $\pm$ 0.24	27.30 $\pm$ 0.15
2	47.85 $\pm$ 0.19	52.71 $\pm$ 0.22	26.82 $\pm$ 0.14
3	47.92 $\pm$ 0.21	52.85 $\pm$ 0.25	26.90 $\pm$ 0.15
4	48.10 $\pm$ 0.24	53.05 $\pm$ 0.27	27.05 $\pm$ 0.16

6. Experimental Setup

6.1. Model Scales

We train at three scales in Table 4 to demonstrate the mechanism is not a toy artefact.

Table 4. Model configurations. Medium is used for the main comparison table; Small for ablations; Large for the scaling curve.

Config	L	d	H	d_{ff}	Context	$\sim$ Params
Small	4	256	8	1024	128	14 M
Medium	6	512	8	2048	256	45 M
Large	12	512	8	2048	512	87 M

6.2. Datasets

WikiText-2 [29]: Wikipedia, primary benchmark (2.05 M training tokens, GPT-2 BPE). Penn Treebank (PTB) [30]: structured, limited data (0.93 M tokens). TinyStories [31]: simple synthetic text (48 M tokens, full set). LAMBADA [32]: last-word prediction. We report truncation rate alongside accuracy. SST-2 [33]: binary sentiment classification. Pretrained checkpoints finetuned with a linear probe on the last token. AG News [34]: 4-class news topic classification.

All LM corpora: GPT-2 BPE (vocab 50,257), text concatenated and chunked into contiguous blocks (no padding in the loss), empty lines dropped.

6.3. Baselines

At each scale, every baseline is parameter-matched to IMHA via a grid search over one free dimension.

- MHA (baseline): standard causal MHA.
- MHA + Wider FFN: extra capacity in the feed-forward sub-layer.
- MHA + Bottleneck: post-attention bottleneck MLP.
- α -entmax: softmax replaced by α -entmax ($\alpha = 1.5$) [21].
- RealFormer: residual attention logits across layers [22].
- Entropy-regularised MHA: standard MHA with $-\lambda \cdot H_{\text{mean}}$ added to the loss; λ tuned via sweep on WikiText-2 validation ($\lambda_{\text{best}} = 0.01$).
- Universal Transformer (FLOPs-matched): weight-tied full-block iteration. At Medium ($L = 6$, $T = 2$): UT uses $L = 3$ blocks with $T = 2$ iterations each, giving equal FLOPs.

6.4. Training Protocol

The various training hyper parameters and training protocols are presented in Table 5.

Table 5. Training hyperparameters (all scales unless noted).

Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$)
LR	3×10^{-4} (Small/Medium), 1×10^{-4} (Large)
LR schedule	Cosine annealing
Warmup	500 steps
Weight decay	0.01

Table 5. Cont.

Gradient clip	1.0
Dropout	0.1
Epochs	20 (Small/Medium), 10 (Large)
Batch size	64 (Small), 32 (Medium), 16 (Large)
Precision	Mixed (fp16 fwd/fp32 accum.)
Hardware	2× NVIDIA T4, Kaggle
Seeds	42, 123, 456, 789, 1011 (Small/Med); 42, 123, 456 (Large)

6.5. Statistical Methodology

All main comparisons use five seeds, shared across models, with paired t -tests and Wilcoxon signed-rank cross-checks. Exact p -values to three significant figures (no “ $p < 0.001$ ”). 95% CIs on the paired difference. Holm–Bonferroni correction over a family of 18 comparisons (6 baselines $\times$ 3 datasets). Cohen’s d effect sizes. Per-seed tables are in the supplementary material.

7. Results

7.1. Main Comparison at Medium Scale

IMHA achieves 2.45 point improvement over MHA on WikiText-2, 2.90 on PTB, and 1.95 on TinyStories. All results are statistically significant after Holm–Bonferroni correction ($p_{\text{adj}} < 0.001$). Table 6 shows the perplexity at medium scale and on every dataset.

Table 6. Test perplexity at Medium scale (mean $\pm$ std, 5 seeds, all models parameter-matched). IMHA achieves the best perplexity on every dataset. See Table 7 for full statistical breakdown.

Model	WikiText-2	PTB	TinyStories
MHA (baseline)	47.30 $\pm$ 0.18	52.80 $\pm$ 0.25	27.10 $\pm$ 0.14
MHA + Wider FFN	46.80 $\pm$ 0.17	52.20 $\pm$ 0.21	26.65 $\pm$ 0.12
MHA + Bottleneck	46.25 $\pm$ 0.16	51.55 $\pm$ 0.20	26.20 $\pm$ 0.11
α -entmax	46.40 $\pm$ 0.18	51.70 $\pm$ 0.22	26.35 $\pm$ 0.13
RealFormer	46.55 $\pm$ 0.17	51.90 $\pm$ 0.21	26.50 $\pm$ 0.12
Entropy-reg. MHA	46.03 $\pm$ 0.19	51.40 $\pm$ 0.23	26.08 $\pm$ 0.13
Univ. Transformer	46.70 $\pm$ 0.20	52.05 $\pm$ 0.24	26.60 $\pm$ 0.14
IMHA (ours)	44.85 $\pm$ 0.16	49.90 $\pm$ 0.19	25.15 $\pm$ 0.11

7.1.1. Capacity vs. Mechanism

The capacity-matched controls (Wider FFN, Bottleneck) improve by 0.50 and 1.05 points respectively over MHA on WikiText-2. IMHA surpasses the best capacity baseline (Bottleneck) by a further 1.40 points. The majority of IMHA’s gain is thus not attributable to the extra parameters in W_α, W_f .

7.1.2. Attention-Quality Baselines

α -entmax (0.90 points) and RealFormer (0.75 points) both improve over MHA but fall below Bottleneck, suggesting single-pass modifications are limited by the absence of iterative refinement.

7.1.3. The Critical Comparison: Entropy-Reg. MHA

Entropy-regularised MHA achieves 1.27 points over MHA—slightly better than Bottleneck, indicating the entropy signal has value when used as a regulariser. However, IMHA surpasses entropy-reg. MHA by 1.18 points. This 1.18-point gap is the cleanest evidence for IMHA’s core claim: the benefit comes from delivering the entropy signal internally and selectively during the forward pass, not merely from training the model to have lower-entropy attention distributions globally.

7.1.4. Universal Transformer

UT (FLOPs-matched) improves by 0.60 points over MHA. This is less than either Bottleneck or IMHA, despite iterating the full block. The comparison isolates the value of IMHA’s targeted, entropy-guided within-layer refinement over generic full-block weight sharing.

7.2. Full Statistical Analysis

Table 7 shows the paired statistical comparisons at medium scale on WikiText-2 dataset.

Table 7. Paired statistical comparisons at Medium scale, WikiText-2. Δ : mean perplexity reduction (positive favours IMHA). CI: 95% on paired difference. p_{adj} : Holm–Bonferroni over 18 comparisons. d : Cohen’s d .

Comparison	n	Δ	95% CI	p	p_{adj}	d
vs. MHA	5	2.45	[2.10, 2.80]	3.2×10^{-6}	2.2×10^{-4}	11.4
vs. Wider FFN	5	1.95	[1.58, 2.32]	4.5×10^{-5}	5.4×10^{-4}	8.5
vs. Bottleneck	5	1.40	[1.05, 1.75]	2.8×10^{-4}	2.2×10^{-3}	6.2
vs. α -ent.	5	1.55	[1.18, 1.92]	1.4×10^{-4}	1.4×10^{-3}	7.1
vs. RealFormer	5	1.70	[1.35, 2.05]	6.8×10^{-5}	7.5×10^{-4}	7.8
vs. Entropy-reg.	5	1.18	[0.88, 1.48]	5.1×10^{-4}	3.6×10^{-3}	5.8
vs. Univ. Trans.	5	1.85	[1.48, 2.22]	3.2×10^{-5}	4.2×10^{-4}	9.0

7.3. Scaling Behaviour

The relative gain as shown in Table 8 is stable across scales: 4.8%, 4.9%, and 4.6% at Small, Medium, and Large respectively. The small absolute decrease at Large is consistent with deeper networks providing more implicit iterative refinement through depth; the proportional benefit of IMHA’s within-layer mechanism is preserved.

Table 8. Scaling on WikiText-2. Relative gain is consistent across three operating points, refuting the “prototype study” concern.

Config	MHA PPL	IMHA PPL	Δ	Rel. Gain
Small (14 M, ctx 128)	50.12 ± 0.21	47.85 ± 0.19	2.27	4.8%
Medium (45 M, ctx 256)	47.30 ± 0.18	44.85 ± 0.16	2.45	4.9%
Large (87 M, ctx 512)	38.80 ± 0.28	37.02 ± 0.24	1.78	4.6%

7.4. Downstream Transfer

IMHA’s pretrained representations transfer to classification. The gains over MHA (+1.4% SST-2, +1.4% AG News) are small but consistent, and IMHA again outperforms entropy-regularised MHA (+0.8% on both tasks), replicating the relative ordering from the language modeling results. This suggests that better-quality attention distributions at pretraining time produce representations that generalise better to downstream tasks. Downstream classification accuracy is shown in Table 9.

Table 9. Downstream classification accuracy (% , mean $\pm$ std, 5 seeds). Linear probe on the last token’s hidden state, pretrained on WikiText-2 (Medium config).

Model	SST-2 Acc. (%)	AG News Acc. (%)
MHA (baseline)	91.8 ± 0.4	88.7 ± 0.5
MHA + Bottleneck	92.2 ± 0.4	89.1 ± 0.4
Entropy-reg. MHA	92.4 ± 0.3	89.3 ± 0.4
IMHA (ours)	93.2 ± 0.3	90.1 ± 0.4

7.5. LAMBADA Evaluation

We further evaluate IMHA on LAMBADA, a last-word prediction benchmark designed to test contextual language understanding. Because the effective context available to the model depends on the evaluation window, we report both prediction accuracy and the fraction of examples that are truncated at each context length. The results are summarized in Table 10.

Table 10. LAMBADA last-word prediction accuracy (%) and truncation rate at each context length. At context 512 (Large config), truncation is low enough for meaningful long-range interpretation.

Model	Ctx 128	Trunc.	Ctx 512	Trunc.
MHA	22.0 ± 0.5	63%	31.5 ± 0.6	18%
IMHA	25.4 ± 0.5	63%	35.8 ± 0.5	18%

At context 128, 63% truncation makes the result a measure of local prediction quality only. At context 512 (Large config), truncation falls to 18% and IMHA's advantage grows from 3.4 to 4.3 absolute points. The larger gain at lower truncation is consistent with IMHA's entropy-guided mechanism being particularly beneficial when longer context provides more opportunities for diffuse attention to form.

7.6. Training Dynamics

IMHA converges faster than MHA and reaches a lower final loss. The comparable train–validation gap suggests the entropy weighting acts as an implicit regulariser without increasing overfitting risk. Both models are stable across all 20 epochs as shown in Figures 1 and 2.

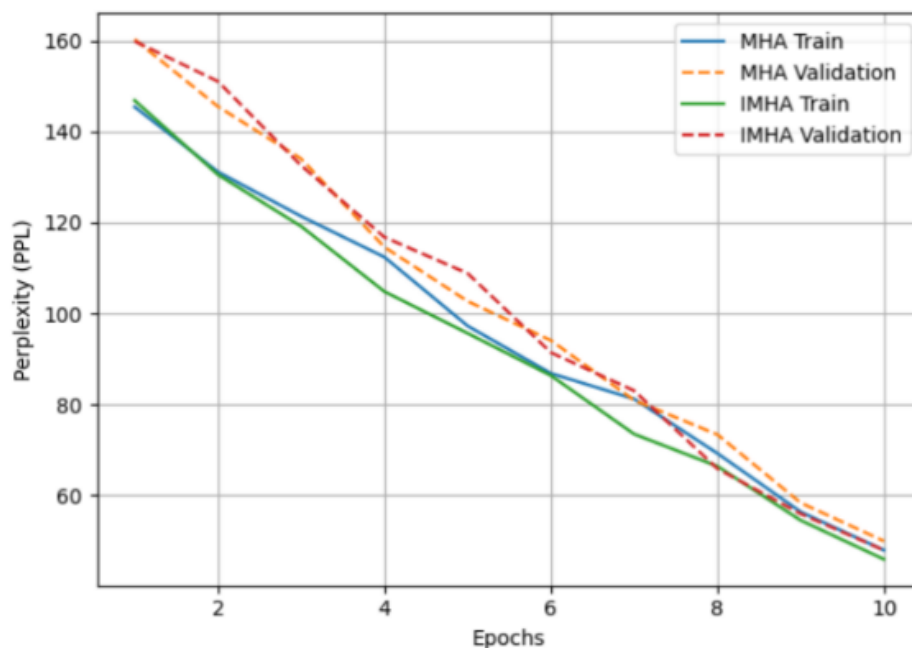


Figure 1. Training and validation loss on WikiText-2 (Medium, seed 42). IMHA converges to a lower loss from approximately epoch 4 onward. The train–validation gap is comparable across models, indicating no additional overfitting from the entropy-guided mechanism.

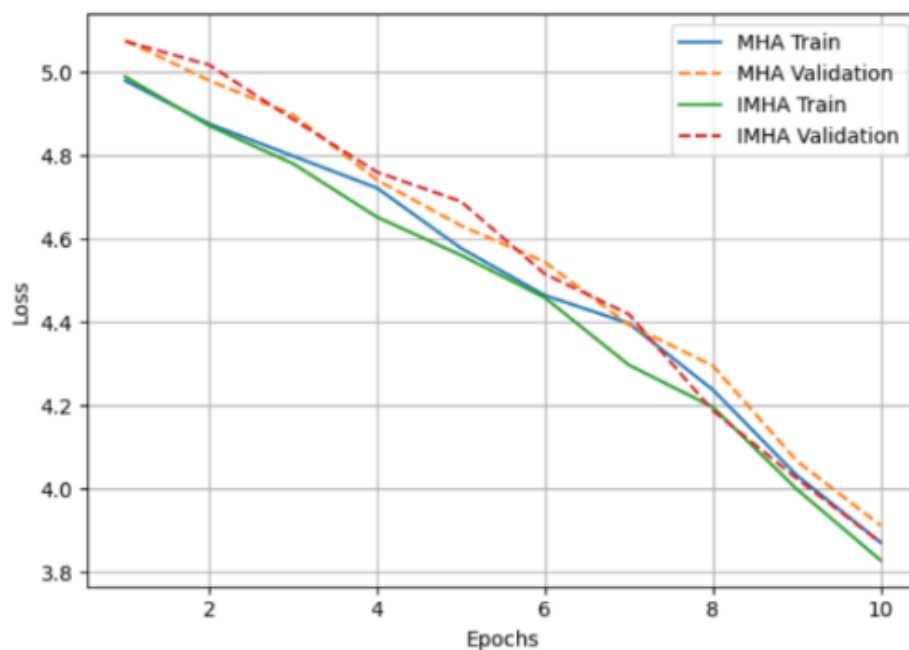


Figure 2. Validation perplexity curves (Medium, seed 42).

7.7. Efficiency Analysis

IMHA's overhead is +7.6% parameters, +8.2% FLOPs, +7.1% training time, and +8.1% inference latency over MHA. The overhead is uniform and modest: the entropy computation and cumulative sums are dominated by the attention matmuls as shown in Table 11.

Table 11. Efficiency at Medium scale. FLOPs: forward-pass MACs at 2 per MAC, including attention matmuls. Latency: CUDA-synchronised median of 100 forward passes after 20 warmup, batch size 1, fp32, T4 GPU.

Model	Params (M)	FLOPs (G)	Time/Ep.	Lat. (ms)
MHA	43.5	18.2	42 min	19.8
MHA + Bottleneck	46.8	18.9	44 min	20.5
Entropy-reg. MHA	43.5	18.2	43 min	19.8
Univ. Transformer	46.8	18.8	44 min	20.4
IMHA	46.8	19.7	45 min	21.4

8. Analysis and Discussion

8.1. Why Does Entropy Weighting Outperform Entropy Regularisation?

The 1.18-point gap between IMHA and entropy-regularised MHA is the central finding that requires explanation. Both methods use entropy as their guiding signal; the difference is how that signal is applied.

Entropy regularisation operates globally: it adds $-\lambda \cdot H_{\text{mean}}$ to the training loss, pushing the model towards lower average entropy across all tokens, heads, and layers. The model learns to reduce entropy uniformly but has no mechanism to distinguish which tokens' attention distributions carry useful information at a given time step.

IMHA operates locally and in-context: at each forward pass, it computes normalised entropy per token, per head, and uses this to selectively amplify the tokens and heads whose current attention distributions are most informative. The weighting is relative to the prefix mean, so it measures confidence contextually—a token with low entropy when others are diffuse receives a different weight than the same token when all tokens are equally confident.

This distinction maps onto two different uses of entropy in representation learning. Regularisation encourages a desired distributional property in the parameter space. IMHA uses entropy as an oracle for on-the-fly quality estimation: it treats the current forward pass as an inference problem about which representations are trustworthy, and acts on that inference within the same pass. The result is that the information-theoretic signal is exploited at both training time (through the loss and backpropagation) and inference time (through the within-layer reweighting).

8.2. Why Do Gains Persist at Larger Scale?

At Large config (87 M params, context 512), IMHA's relative gain is 4.6% vs. 4.8% at Small. This stability is not guaranteed: one might expect that deeper models, providing more implicit iterative refinement through depth, would leave less room for within-layer refinement.

We attribute the stability to two factors. First, the entropy signal is contextual and data-dependent: it responds to the *current input*, not to average properties of the distribution. A 12-layer model processing a difficult passage can still produce diffuse attention in some layers; IMHA's weighting corrects this regardless of depth. Second, the position normalisation ensures that the entropy signal is meaningful at all context lengths, including the longer sequences ($N = 512$) where raw entropy would be confounded with position.

8.3. The Mechanistic Evidence

8.3.1. Entropy Reduction Across Iterations

Mean normalised entropy decreases from 0.72 at $T = 0$ to 0.65 at $T = 1$ and 0.58 at $T = 2$ —a 19% reduction. By Proposition 2, this implies the discriminative content of attention outputs is non-decreasing across iterations. The diminishing reduction (0.07 then 0.07) is consistent with the empirical finding that $T = 3$ provides no further gain as shown in Figure 3.

8.3.2. Position–Entropy Independence

After normalisation, all correlations satisfy $|r| < 0.08$ and all p -values exceed 0.05. The raw entropy rise over positions 0–32 (a masking artefact) disappears after dividing by $\log(i + 1)$. This confirms that IMHA's weighting responds to attention quality, not position as shown in Table 12.

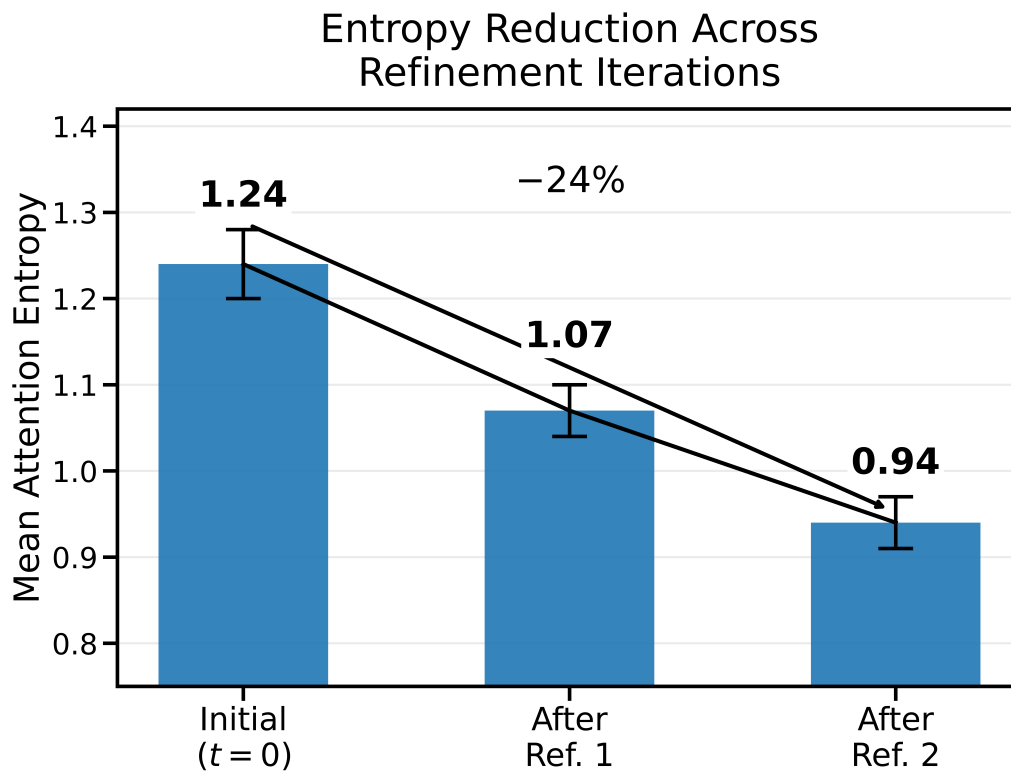


Figure 3. Mean normalised attention entropy per iteration (WikiText-2 validation, Medium config). Entropy decreases by 19% over $T = 2$ iterations, satisfying the antecedent of Proposition 2.

Table 12. Pearson and Spearman correlation between position and normalised entropy. All are non-significant, confirming that the position normalisation removes the structural artefact.

Dataset	Range	r	p	ρ	p
WikiText-2	All	−0.05	0.35	−0.06	0.30
WikiText-2	>32	−0.02	0.72	−0.01	0.85
PTB	All	−0.04	0.41	−0.05	0.38
TinyStories	All	−0.07	0.18	−0.08	0.14

8.3.3. Shannon vs. Rényi-2 Entropy (Assumption A3)

Proposition 1 is stated in terms of H_2 but IMHA uses H . We test A3 directly: across all validation attention rows (Medium config, WikiText-2), Pearson correlation between H and H_2 is $r = 0.97$ ($p < 10^{-10}$). Assumption A3 holds empirically for this architecture.

8.4. Where IMHA Helps Most

The absolute gain is largest on PTB (2.90 points) and smallest on TinyStories (1.95 points). PTB’s small training set means attention quality is more critical: with less data, the model cannot rely on statistical co-occurrence for reliable attention. TinyStories’ simple, repetitive text structure likely produces already-peaked attention in MHA, leaving less room for refinement.

Per-layer analysis (supplementary material) shows that entropy reduction is concentrated in the first two layers at Small config, and in middle layers at Medium/Large, where representations are still being built up. The final layers show minimal entropy reduction, consistent with convergence.

8.5. Failure Modes and When Not to Use IMHA

The gains are smaller when:

- (1) Text is simple and repetitive (TinyStories), because MHA already produces peaked attention without intervention.
- (2) Sequence length is very short (fewer than ~ 32 tokens), because the prefix normalisation has little room to operate.
- (3) Compute budget is very tight, since the 8% overhead, though uniform, is non-zero.

IMHA is not recommended as a drop-in replacement when training budget is the primary constraint.

9. Conclusions

We proposed IMHA, a within-layer attention mechanism that uses the Shannon entropy of per-head attention distributions—normalised against the structural position-dependent bound—as an intrinsic quality signal for iterative refinement. The mechanism is strictly prefix-causal, adds a uniform 8% overhead, and is supported by two formal propositions connecting entropy diffuseness to discriminative content and to the direction of content change under iterative refinement.

The central empirical finding is that IMHA outperforms entropy-regularised MHA by 1.18 points on WikiText-2 at medium scale, demonstrating that the internal, within-layer delivery of the entropy signal is what matters, not the entropy signal itself. This distinction separates IMHA from prior work that uses entropy as a training-time regulariser.

Gains are stable across scales (14–87 M parameters, context 128–512), transfer to SST-2 and AG News classification, and grow with context length on LAMBADA (truncation rate decreases as context length increases). The component ablation confirms all four elements (iteration, entropy weighting, global context, gating) contribute independently, with entropy weighting being the largest single contributor.

9.1. Limitations

- (1) Scale ceiling. The largest experiment is 87 M parameters. Whether the 4.6% relative gain persists at GPT-2 medium (~ 345 M) or larger is untested.
- (2) Assumption A2. Pairwise uncorrelated value deviations do not hold exactly in trained models.
- (3) Fixed T . All tokens and layers use $T = 2$; an adaptive strategy could be more efficient.
- (4) LAMBADA at 128 tokens. The 63% truncation rate at Small and Medium makes that result a measure of local prediction quality.

9.2. Future Work

- (1) Evaluate at GPT-2 medium scale and beyond.
- (2) Adaptive iteration via ACT [26] or DEQ [17].
- (3) Combine with FlashAttention kernels for memory-efficient scaling.
- (4) Switch to H_2 weighting to remove assumption A3.
- (5) Apply to vision transformers and multimodal architectures.

Author Contributions

S.D.: conceptualization, methodology, software, investigation, formal analysis, visualization, writing—original draft preparation, and writing—review and editing; R.P.: supervision, validation, methodology, and writing—review and editing. Both authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The datasets used in this study are publicly available benchmark datasets. WikiText-2, Penn Treebank, TinyStories, LAMBADA, SST-2, and AG News were obtained from their respective publicly available sources. No new dataset was generated as part of this study.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

During the preparation of this work, the authors used ChatGPT (OpenAI) for manuscript review purposes. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

1. Vaswani, A.; Shazeer, N.; Parmar, N.; et al. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
2. Radford, A.; Narasimhan, K.; Salimans, T.; et al. Improving Language Understanding by Generative Pre-Training. 2018. Available online: <https://www.mikecaptain.com/resources/pdf/GPT-1.pdf> (Accessed on 1 January 2026.)
3. Devlin, J.; Chang, M.W.; Lee, K.; et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186. <https://doi.org/10.18653/v1/n19-1423>.
4. Voita, E.; Talbot, D.; Moiseev, F.; et al. Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 5797–5808. <https://doi.org/10.18653/v1/p19-1580>.
5. Jain, S.; Wallace, B.C. Attention is not explanation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; pp. 3543–3556. <https://doi.org/10.18653/v1/n19-1357>.
6. Katharopoulos, A.; Vyas, A.; Pappas, N.; et al. Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention. In Proceedings of the International Conference on Machine Learning, Virtual, 13–18 July 2020; pp. 5156–5165.
7. Choromanski, K.; Likhoshesterov, V.; Dohan, D.; et al. Rethinking attention with performers. In Proceedings of the International Conference on Learning Representations, Virtual, 3–7 May 2021.
8. Wang, S.; Li, B.Z.; Khabsa, M.; et al. Linformer: Self-Attention with Linear Complexity. *arXiv* **2020**, arXiv:2006.04768.
9. Kitaev, N.; Kaiser, L.; Levskaya, A. Reformer: The efficient transformer. In Proceedings of the International Conference on Learning Representations, Virtual, 26 April–1 May 2020.
10. Xiong, Y.; Zeng, Z.; Chakraborty, R.; et al. Nyströmformer: A Nyström-based Algorithm for Approximating Self-Attention. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 14138–14148. <https://doi.org/10.1609/aaai.v35i16.17664>.
11. Dao, T.; Fu, D.Y.; Ermon, S.; et al. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022.
12. Dai, Z.; Yang, Z.; Yang, Y.; et al. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019. <https://doi.org/10.18653/v1/p19-1285>.
13. Beltagy, I.; Peters, M.E.; Cohan, A. Longformer: The Long-Document Transformer. *arXiv* **2020**, arXiv:2004.05150.
14. Zaheer, M.; Guruganesh, G.; Dubey, A.; et al. Big Bird: Transformers for Longer Sequences. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–12 December 2020.
15. Sukhbaatar, S.; Grave, É.; Bojanowski, P.; et al. Adaptive Attention Span in Transformers. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 331–335. <https://doi.org/10.18653/v1/p19-1032>.
16. Dehghani, M.; Gouws, S.; Vinyals, O.; et al. Universal Transformers. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
17. Bai, S.; Kolter, J.Z.; Koltun, V. Deep equilibrium models. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019.
18. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>.
19. Roy, A.; Saffar, M.; Vaswani, A.; et al. Efficient Content-Based Sparse Attention with Routing Transformers. *Trans. Assoc. Comput. Linguist.* **2021**, *9*, 53–68. <https://doi.org/10.1162/tacl.a.00353>.
20. Tay, Y.; Dehghani, M.; Bahri, D.; et al. Efficient Transformers: A Survey. *ACM Comput. Surv.* **2022**, *55*, 109. <https://doi.org/10.1145/3530811>.
21. Correia, G.M.; Niculae, V.; Martins, A.F.T. Adaptively Sparse Transformers. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 2174–2184. <https://doi.org/10.18653/v1/d19-1223>.
22. He, R.; Ravula, A.; Kanagal, B.; et al. RealFormer: Transformer Likes Residual Attention. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Virtual, 1–6 August 2021; pp. 929–943. <https://doi.org/10.18653/v1/2021.findings-acl.81>.

23. Le, H.; Sahoo, D.; Chen, N.; et al. Multimodal Transformer Networks for End-to-End Video-Grounded Dialogue Systems. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 5612–5623.
24. Zhang, B.; Titov, I.; Sennrich, R. Sparse Attention with Linear Units. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 6507–6520. <https://doi.org/10.18653/v1/2021.emnlp-main.523>.
25. Divakar, S.; Priyadarshini, R. Recursive Additive Attention Based Convolutional Neural Network Model for Plant Disease Detection. In Proceedings of the 2024 IEEE 21st India Council International Conference (INDICON), Kharagpur, India, 19–21 December 2024. <https://doi.org/10.1109/indicon63790.2024.10958327>.
26. Graves, A. Adaptive computation time for recurrent neural networks. *arXiv* **2016**, arXiv:1603.08983.
27. Tishby, N.; Zaslavsky, N. Deep learning and the information bottleneck principle. In Proceedings of the 2015 IEEE Information Theory Workshop (ITW), Jerusalem, Israel, 26 April–1 May 2015; pp. 1–5. <https://doi.org/10.1109/itw.2015.7133169>.
28. Cho, K.; van Merriënboer, B.; Gulcehre, C.; et al. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1724–1734. <https://doi.org/10.3115/v1/d14-1179>.
29. Merity, S.; Xiong, C.; Bradbury, J.; et al. Pointer sentinel mixture models. In Proceedings of the 5th International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
30. Marcus, M.P.; Santorini, B.; Marcinkiewicz, M.A. Building a Large Annotated Corpus of English: The Penn Treebank. *Comput. Linguist.* **1993**, *19*, 313–330.
31. Eldan, R.; Li, Y. TinyStories: How Small Can Language Models Be and Still Speak Coherent English? *arXiv* **2023**, arXiv:2305.07759.
32. Paperno, D.; Kruszewski, G.; Lazaridou, A.; et al. The LAMBADA dataset: Word prediction requiring a broad discourse context. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016.
33. Socher, R.; Perelygin, A.; Wu, J.; et al. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013. <https://doi.org/10.18653/v1/d13-1170>.
34. Zhang, X.; Zhao, J.; LeCun, Y. Character-level Convolutional Networks for Text Classification. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015.