Entropy Meets Importance: A Unified Head Importance–Entropy Score for Stable and Efficient Transformer Pruning

Minsik Choi* MSZZANG2002@KOREA.AC.KR *Korea University, Seoul, Republic of Korea*
Hyegang Son* HYEGANG_SON@KOREA.AC.KR *Korea University, Seoul, Republic of Korea*
Joohun Hyun* JOOHUNHYUN@KOREA.AC.KR *Korea University, Seoul, Republic of Korea*
Seokmin Kim* WINTER02@KOREA.AC.KR *Korea University, Seoul, Republic of Korea*
Young Geun Kim† YOUNGGEUN_KIM@KOREA.AC.KR *Korea University, Seoul, Republic of Korea*

Abstract

Transformer-based models have achieved remarkable performance in NLP tasks. However, their structural characteristics—multiple layers and attention heads—introduce challenges in inference and deployment. To address these challenges, various pruning methods have recently been proposed. Notably, gradient-based methods using Head Importance Scores (HIS) have gained traction for interpretability, efficiency, and ability to identify redundant heads. However, HIS alone has limitations as it captures only the gradient-driven contribution, overlooking the diversity of attention patterns. To overcome these limitations, we introduce a novel pruning criterion, **HIES (Head Importance-Entropy Score)**, which integrates head importance scores with attention entropy, providing complementary evidence on per-head contribution. Empirically, HIES-based pruning yields up to 17.62% improvement in model quality and 2.05× improvement in stability over HIS-only methods, enabling substantial model compression without sacrificing either accuracy or stability.

Keywords: List of keywords: Transformer Architecture, Attention Head Pruning, Model Stability, Attention Entropy

1. Introduction

Recent advances in Large Language Models (LLMs) have led to remarkable performance. In pursuit of better modeling of long-range dependencies, LLMs have scaled up both context lengths and attention head counts, guided by empirical scaling laws that correlate model capacity with performance [9]. This however substantially increases computational and memory costs during inference, which in turn leads to higher latency and energy consumption. These issues are particularly problematic when LLMs are deployed to resource-constrained environments such as consumer-grade mobile devices or edge devices, for applications including real-time translation, intelligent voice assitants and personalized recommendation systems.

To improve deployability of LLMs for resource-constrained environments, various pruning methods have been proposed. In typical, these methods selectively reduce computations by removing less important weights, channels, or attention heads. Among them, *head pruning* has gained considerable attention due to its structural simplicity, interpretability, and ability to directly target redundancy within the attention mechanism. Existing head pruning methods typically identifies less

. * Equal contribution. † Corresponding author.

important heads based on Head Importance Score (HIS), which measures the gradient-based contribution of each head to the loss function. By leveraging gradient-based sensitivity to the loss, HIS prioritizes components that have the most direct impact on model behavior during inference.

However, HIS-based methods often exhibit limited stability in their performance. For clarity, prior works [1, 2] motivate treating stability as a practical surrogate for robustness—namely, a model's resilience to input perturbations and pruning-induced changes. Such stability is crucial in real-world deployments where distribution shifts are common and aggressive compression is often required. Consequently, this limited stability can be attributed to two key factors. First, existing HIS-based methods solely rely on the loss gradient with respect to each head's output, which does not reflect token-level attention allocation or its alignment with the task's empirical distribution. Hence, a concentrated head and a diffuse head can appear equivalent, concealing their functionally different behavior on the task-specific data manifold. Second, a uniform, layer-agnostic criterion precludes layer-specific adaptation, since different heads often require distinct attention behaviors. Lacking such layer-specific characteristics often results in imbalanced pruning—preserving redundant heads in some layers while removing functionally important ones in others. This imbalance not only degrades accuracy but also undermines stability, leading to unpredictable performance fluctuations across inputs or compression levels, especially under high pruning ratios.

This work aims to address the aforementioned limitations by proposing an *Entropy-Aware Pruning Criterion*, termed **HIES(Head Importance-Entropy Score)**, considers a head's gradient-based contribution to the loss and the distributional structure of its attention—specifically, the extent to which its attention is concentrated or dispersed across input tokens. Specifically, we compute the HIS to quantify a head's loss relevance and Attention Entropy (AE) to measure how evenly a head distributes attention over input tokens. Their combination, HIES, enables layer-sensitive pruning decisions and mitigates the risk of removing functionally important heads. This allows for more balanced pruning across layers, which in turn improves both accuracy and stability under aggressive compression. Empirically, HIES yields up to a 17.62% improvement in model quality and 2.05× improvement in stability over HIS-only methods. Given its ability to preserve both accuracy and stability even under aggressive pruning ratios, HIES represents a more practical and robust solution compared to existing pruning methods. It is expected to offer more stable performance in resource-constrained environments.

2. Background and Related Work

Attention Head Pruning. To compress large language models efficiently, structured pruning methods, which remove specific architectural components from Transformer models, have been widely adopted. Among these, attention head pruning has gained traction. This is largely because it directly lowers attention FLOPs and KV-cache by reducing the number of heads while preserving the layer topology, thereby simplifying checkpoint compatibility and serving integration. Consequently, large-scale studies adopt head-level pruning as a practical axis in LLM compression pipelines [8, 13]. Attention head pruning removes selected heads from a trained Transformer's multi-head attention with minimal impact on end-task performance [15]. A widely used criterion is the HIS of Michel et al. [12], which introduces a binary gate $m_h \in \{0, 1\}$ multiplying the output of head h and defines importance as the expected first-order loss increase under masking:

$$\mathrm{HIS}_h = \mathbb{E}_{x\sim\mathcal{D}} \left| \frac{\partial \mathcal{L}(x)}{\partial m_h} \right| = \mathbb{E}_{x\sim\mathcal{D}} \left| \mathrm{A}_h(x)^\top \frac{\partial \mathcal{L}(x)}{\partial \mathrm{A}_h(x)} \right| . \tag{1}$$

Here, $\mathcal{D}$ denotes the data distribution, $\mathcal{L}(x)$ is the loss for sample x, and $\mathrm{A}_h(x)$ is the output of head h. The second equality follows from the chain rule and the observation that gating scales the head's activation. Heads are then ranked by I_h and pruned in ascending order of importance.

Attention Entropy and Stability. Zhai et al. [19] quantifies the concentration of each attention head's focus over input tokens via the entropy of its attention weight distribution $\mathrm{AE}_h = (H\big(p^{(h)}\big) = -\sum_{i=1}^{n} p_i^{(h)} \log p_i^{(h)}$, where $p_i^{(h)}$ is the normalized attention probability assigned by head h to the i-th input token subject to $\sum_{i=1}^{n} p_i^{(h)} = 1$. Higher entropy indicates a diffuse focus over the sequence, whereas lower entropy corresponds to highly concentrated attention patterns. Their empirical findings reveal a strong correlation between persistently low entropy (i.e. entropy collapse) and instability during training, such as oscillations in the loss landscape or even divergence across various model scales and tasks.

We hypothesize that attention entropy may also reflect aspects of inference-time instability. In particular, low entropy may correlate with increased sensitivity to input perturbations, degraded calibration, or inconsistent performance under resource constraints. While the underlying causal relationship between attention entropy and training stability remains unclear, we posit that attention entropy captures structural signals that may contribute to or reflect unstable behavior, both during training and deployment. This motivates our investigation of entropy as a proxy for robustness and consistency during inference.

3. Proposed Method

3.1. Head Importance Entropy Score

We define the **Head Importance–Entropy Score (HIES)** as a weighted combination:

$$\mathrm{HIES}_h = \alpha\, \widehat{\mathrm{HIS}}_h + (1-\alpha)(1-\widehat{\mathrm{AE}}_h), \quad \alpha \in [0, 1), \tag{1}$$

where α[1] is a tunable hyperparameter.

Min-Max Normalization Directly comparing raw HIS and AE is inherently problematic, as the two metrics reside on different scales and encode distinct types of signal. To enable meaningful integration and ranking, we apply *min–max normalization* to both metrics, rescaling their values to the interval $[0, 1]$: $\widehat{\mathrm{HIS}}_h = \frac{\mathrm{HIS}_h - \min(\mathrm{HIS})}{\max(\mathrm{HIS}) - \min(\mathrm{HIS})}$, $\widehat{\mathrm{AE}}_h = \frac{\mathrm{AE}_h - \min(\mathrm{AE})}{\max(\mathrm{AE}) - \min(\mathrm{AE})}$. This distribution-agnostic normalization improves outlier robustness and cross-criterion interpretability; lower normalized scores denote higher pruning priority. Prior studies show min–max scaling outperforms z-score standardization in stability and reproducibility across diverse tasks [5, 10].

3.2. Theoretical Analysis

We analyze pruning through a risk decomposition that combines a loss-increase term controlled by HIS, with a generalization-gap term upper-bounded in terms of AE via its tokenwise deficit. We further show that the gradients of HIS and AE are orthogonal in expectation, indicating complementary axes: magnitude of contribution (HIS) and dispersion of attention (AE). This perspective motivates the composite importance measure HIES. By keeping heads with high HIES, we simultaneously minimize our theoretical bound and enhance pruning stability. Conceptually, this analysis

[1] To determine the optimal combination of HIS and AE for each task, we adopt a task-specific tuning procedure based on weighted AUC (wAUC), which selects the best-performing combination under each compression setting. Sensitivity to α is analyzed in Appendix C.3.

formalizes importance-based selection into a principled framework and offers a rigorous rationale for HIES's safety and effectiveness.

3.2.1. LOSS-INCREASE CONTROL VIA HEAD IMPORTANCE (HIS)

Setup. Let n be the sequence length, H the number of heads, d_v the value dimension per head, and $d = H\, d_v$ the model width (i.e., hidden size). For head h, let $A_h \in \mathbb{R}^{n\times d_v}$ be its output and define $\mathbf{y} = \mathrm{Concat}(A_1, \ldots, A_H) \in \mathbb{R}^{n\times d}$ as the pre-projection representation, projected through $W^O \in \mathbb{R}^{d\times d}$. Head removal is modeled by binary gates $m_h \in \{0,1\}$: $\delta A_h = -(1-m_h)A_h$ and $\delta\mathbf{y} = \mathrm{Concat}(\delta A_1, \ldots, \delta A_H)$. Formal definitions and implementation notes are deferred to Appendix A.1.2.

Head Importance Score (HIS). We define

$$\mathrm{HIS}_h \;:=\; \mathbb{E}_{x\sim\mathcal{D}}\left|\frac{\partial\mathcal{L}(x)}{\partial\xi_h}\right| \;=\; \mathbb{E}_{x\sim\mathcal{D}}\big|\,\big\langle\nabla_{A_h(x)}\mathcal{L}(x),\, A_h(x)\big\rangle_F\,\big| \;=\; \mathbb{E}_{\mathcal{D}}\Big[\,|\,\langle\nabla_{A_h}\mathcal{L},\, A_h\rangle_F\,|\,\Big]. \quad (2)$$

This quantity is a first-order activation–gradient correlation whose absolute value prevents cross-sample cancellation, yielding the additive upper bound $\sum_h \mathrm{HIS}_h$ on the loss (cf. Appendix A.1.6).

Theorem 1 (Loss-increase upper bound under head masking) *Let $\beta_y := \|\nabla^2_{\mathbf{y}}\mathcal{L}\|_2$. For any binary gates $\{m_h\}_{h=1}^{H}$,*

$$\Delta\mathcal{L} \;:=\; \mathbb{E}_{\mathcal{D}}\big[\mathcal{L}(\mathbf{y}+\delta\mathbf{y}) - \mathcal{L}(\mathbf{y})\big] \;\le\; \sum_{h=1}^{H}(1-m_h)\,\mathrm{HIS}_h \;+\; \frac{\beta_y}{2}\sum_{h=1}^{H}(1-m_h)\,\|A_h\|_F^2. \quad (3)$$

Moreover, under binary (sigmoid) cross-entropy *we have $\|\nabla^2_{\mathbf{z}}\mathcal{L}\|_2 \le \frac{1}{4}$; with the linear projection $\mathbf{z} = \mathbf{y}W^O$, this yields $\beta_y \le \frac{1}{4}\,\|W^O\|_2^2$, hence*

$$\boxed{\Delta\mathcal{L} \;\le\; \sum_{h=1}^{H}(1-m_h)\,\mathrm{HIS}_h \;+\; \frac{1}{8}\,\|W^O\|_2^2\sum_{h=1}^{H}(1-m_h)\,\|A_h\|_F^2\;.} \quad (4)$$

Remark. For *multiclass softmax* cross-entropy, $\|\nabla^2_{\mathbf{z}}\mathcal{L}\|_2 \le \frac{1}{2}$ (cf. Appendix A.1.4); thus the quadratic coefficient becomes $\frac{1}{4}$ instead of $\frac{1}{8}$.

Implication for pruning. Eq. (4) shows that, for a fixed pruning fraction $\rho = \frac{1}{H}\sum_h(1-m_h)$, selecting heads with the smallest HIS_h minimizes the dominant first-order term, while the quadratic term is controlled by $\|W^O\|_2$ (or blockwise norms) and token-averaged activations. Under standard normalization, the quadratic contribution is typically dominated by the first-order term (cf. Appendix A.1.5), which justifies using HIS_h as a practical surrogate importance for head pruning.

3.2.2. GENERALIZATION GAP AND ATTENTION ENTROPY (AE)

Setup. Let $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{N} \sim \mathcal{D}$. We write $\mathbb{E}_{\mathcal{S}}$ and $\mathbb{E}_{\mathcal{D}}$ for empirical and population expectations. We assume the per-example loss ℓ is L_ℓ-Lipschitz in its first argument, a standard assumption that yields stability-based generalization bounds [3, 7, 14].

Notation (attention entropy deficit). For head h and query token $t \in [n(x)]$, let $\boldsymbol{\alpha}_t^{(h)}(x) \in \Delta^{n(x)-1}$ be the attention over keys and $H(\mathbf{p}) := -\sum_j p_j \log p_j$. Define the token-averaged, length-normalized deficit (AD)

$$\mathrm{AD}_h(x) := \frac{1}{n(x)\,\log n(x)}\sum_{t=1}^{n(x)}\Big(\log n(x) - H\big(\boldsymbol{\alpha}_t^{(h)}(x)\big)\Big) \;=\; 1 - \mathrm{AE}_h(x) \in [0,1].$$

Main bound (loss–entropy link). Let $M := \max_h \max_j \|V_h(j,:)\|_2$ and $C_{\mathrm{AE}} := \sqrt{8}\, M \sqrt{H\rho \log n}$ for a representative effective length n. For the pruned model $f_{\mathcal{S},m}$, the expected generalization gap ($\mathcal{G}$) satisfies

$$\boxed{\begin{aligned} \mathcal{G} &:= \mathbb{E}_{\mathcal{D}}\big[\ell(f_{\mathcal{S},m}(x),y)\big] \;-\; \mathbb{E}_{\mathcal{S}}\big[\ell(f_{\mathcal{S},m}(x),y)\big] \\ &\le\; 2L_\ell\, C_{\mathrm{AE}} \sqrt{\sum_{h=1}^{H}(1-m_h)\,\mathbb{E}_{\mathcal{S}}\big[\overline{\mathrm{AD}}_h(x)\big] \;+\; \frac{B}{N}}\,. \end{aligned}} \tag{5}$$

Interpretation. For a fixed pruning ratio ρ, pruning heads with smaller deficit (i.e., higher entropy) minimizes the bound's increase; pruning low-entropy (high-deficit) heads worsens it. Proof details, operator-norm assumptions, and variable-length handling are deferred to Appendix A.2.

3.2.3. RISK UPPER BOUND AND HIES MINIMIZATION

Composite Risk Bound. Given the HIES defined above, the overall risk upper bound is

$$\mathcal{R}(m) \;:=\; \sum_{h=1}^{H}(1-m_h)\,\mathrm{HIES}_h \tag{1}$$

Pruning Objective (fixed budget). Let $k := (1-\rho)H$ be the number of heads to retain. We solve the cardinality-constrained selection problem

$$\min_{m\in\{0,1\}^H} \sum_{h=1}^{H}(1-m_h)\,\mathrm{HIES}_h \quad \text{s.t.} \sum_{h=1}^{H} m_h \;=\; k. \tag{2}$$

Theorem 2 (Optimality) *Selecting the k heads with the largest HIES values (equivalently, pruning the $H-k$ heads with the smallest HIES values) yields the globally optimal mask m^* that minimizes (1) subject to (2).*

3.2.4. ORTHOGONALITY AND COMPLEMENTARITY

Theorem 3 (Orthogonality) *Let*

$$u_h \;:=\; \mathrm{sign}\big(\boldsymbol{\alpha}^{(h)\top} g_h\big)\, g_h, \qquad v_h \;:=\; \mathbf{1} + \log \boldsymbol{\alpha}^{(h)}, \qquad \tilde{u}_h := P\,u_h,\;\; \tilde{v}_h := P\,v_h,$$

where $P := I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ projects onto $\{w : \mathbf{1}^\top w = 0\}$. Assume $\mathrm{Cov}(\tilde{u}_h, \tilde{v}_h) = 0$ (the cross-covariance matrix is zero). If, in addition, either $\mathbb{E}_{\mathcal{S}}[\tilde{u}_h] = 0$ or more generally $\langle \mathbb{E}_{\mathcal{S}}[\tilde{u}_h], \mathbb{E}_{\mathcal{S}}[\tilde{v}_h]\rangle = 0$, then

$$\mathbb{E}_{\mathcal{S}}\Big[\big\langle \widetilde{\nabla}_{\boldsymbol{\alpha}^{(h)}}\mathrm{HIS}_h,\; \widetilde{\nabla}_{\boldsymbol{\alpha}^{(h)}}\mathrm{AE}_h\big\rangle\Big] \;=\; 0,$$

i.e., the two gradient directions are orthogonal in expectation.[2]

Complementarity. Because the gradients point along statistically orthogonal directions, HIS captures the magnitude of loss sensitivity whereas AE captures the dispersion of attention. Thus, they serve complementary roles: HIS emphasizes the magnitude of contribution, while AE characterizes the distributional concentration. Combined, they balance pruning minimally influential heads and preserving heads important for generalization, underpinning HIES's effectiveness.

[2] Detailed derivations and preliminaries are deferred to Appendix A.4.

Table 1: Experimental results with $\text{BERT}_{\text{base}}$ on natural language understanding task. We report percentage improvements in blue.

Pruning Ratio	Method	SST-2 Accuracy	CoLA Matthews corr	MRPC F1 Score	QQP Accuracy	STS-B Pearson corr	QNLI Accuracy	MNLI Accuracy	RTE Accuracy	Average	
Ratio = 0%	$\text{BERT}_{\text{base}}$	88.76	76.69	91.35	91.27	94.02	91.54	83.75	72.56	86.24	
Ratio = 10%	Random	88.30	75.25	89.90	91.00	93.88	89.87	82.62	70.76	85.20	0.00%
	AD	88.19	75.48	90.56	91.10	93.65	90.87	83.50	70.40	85.47	+0.32%
	HIS	88.88	77.21	91.16	90.63	94.04	90.88	83.00	72.20	**86.00**	**+0.94%**
	HIES (ours)	88.19	74.70	90.84	90.61	93.97	90.70	83.00	71.12	85.39	+0.23%
Ratio = 30%	Random	85.44	69.02	85.27	84.00	92.90	79.80	75.15	60.29	78.98	0.00%
	AD	86.58	50.00	84.53	84.57	81.94	67.82	77.00	56.68	73.64	-6.77%
	HIS	84.06	74.90	89.88	89.95	93.89	88.98	82.25	70.76	84.33	+6.77%
	HIES (ours)	86.58	74.04	88.64	89.62	93.88	89.13	81.25	70.40	**84.19**	**+6.59%**
Ratio = 50%	Random	81.88	61.02	72.53	66.25	91.40	67.53	67.32	53.79	70.22	0.00%
	AD	82.91	50.00	54.50	76.18	75.00	68.94	68.00	55.96	66.44	-5.38%
	HIS	76.03	61.14	86.89	85.91	92.46	81.86	76.50	65.34	78.27	+11.47%
	HIES (ours)	82.91	71.91	87.93	85.37	92.16	82.12	76.50	62.82	**80.22**	**+14.24%**

4. Experimental Results

4.1. Main Results

We evaluate HIES using two key metrics: model quality and stability.[3] As reported in Table 1, HIES improves model quality by 7.02% on average. Table 2 further indicates a 5.4% average stability gain over HIS. Notably, at a pruning ratio of 50%, HIES achieves gains of up to 17.62% in model quality and 2.05× in stability compared to the best-performing baseline. These results corroborate our theoretical analysis and indicate that HIES preserves end-task performance while markedly enhancing robustness, both of which are critical for reliable and efficient deployment.

Table 2: Stability results for $\text{BERT}_{\text{base}}$ on GLUE across pruning ratios.

Pruning Ratio	Method	SST-2	CoLA	MRPC	QQP	Average	
Ratio = 10%	HIS	94.84	96.55	94.85	97.14	**95.85**	**0.00%**
	HIES (ours)	96.22	94.63	93.87	96.92	95.41	-0.45%
Ratio = 20%	HIS	88.99	94.15	94.12	96.48	93.44	0.00%
	HIES (ours)	94.84	94.53	91.18	96.42	**94.24**	**+0.86%**
Ratio = 30%	HIS	85.67	90.60	91.42	94.86	90.64	0.00%
	HIES (ours)	89.79	93.67	87.25	95.04	**91.44**	**+0.88%**
Ratio = 40%	HIS	82.34	72.48	86.52	92.58	83.48	0.00%
	HIES (ours)	87.73	91.95	87.25	92.28	**89.80**	**+7.57%**
Ratio = 50%	HIS	76.26	43.62	84.07	87.78	72.93	0.00%
	HIES (ours)	83.37	89.36	84.56	87.34	**86.16**	**+18.13%**

4.2. Extended Experimental Results

4.2.1. HEAD REMOVAL PATTERNS (HEATMAP)

In main results, HIES exhibits more stable performance gains than HIS at higher pruning ratios (e.g., $\geq$,30%). At lower ratios (e.g., $\leq$,10%), performance is broadly comparable, with HIS sometimes slightly ahead. We posit distinct prioritization. HIS's one-step gradient metric retains redundant, low-risk heads at low sparsity, while HIES adds attention entropy to capture stability, thereby preserving specialized low-entropy heads that improve robustness. This distinction between HIES and HIS is empirically supported by our heatmap analysis in Appendix C.2. Accordingly, HIES exhibits more stable performance across pruning ratios, yielding flatter accuracy–sparsity curves than HIS.

4.2.2. SCALABILITY TO LARGER TRANSFORMER MODELS

We assess performance at scale via zero-shot classification on LLaMA-7B. As shown in Appendix C.4, HIES maintains or improves accuracy and, critically, achieves greater stability than HIS as pruning increases. Crucially, these advantages persist at the 7B scale across the reported pruning ratios (10–60%), underscoring that HIES scales reliably to larger models with higher head counts.

[3] A detailed description of the experimental setup is provided in Appendix B.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIT) (RS-2025-24534857); in part by the Institute of Information & Communications Technology Planning & Evaluation - Information Technology Research Center (IITP-ITRC) Grant funded by MSIT (No. RS-2024-00398353, Development of Countermeasure Technologies for Generative AI Security Threats); Institute of Information & Communications Technology Planning & Evaluation - Information Technology Research Center (IITP-ITRC) grant funded by the MSIT (IITP-2025-RS-2023-00260091); in part by ICT Creative Consilience Program through the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by MSIT (IITP-2025-RS-2020-II201819). We would like to express our sincere gratitude to the members of the Computer Architecture & Systems Research Laboratory at Korea University for their constructive feedback, thoughtful suggestions, and continuous encouragement throughout this work. We also thank the anonymous reviewers for their careful evaluation and insightful comments, which helped improve the clarity of our presentation and the quality of the manuscript.

References

[1] Anna Bair, Hongxu Yin, Maying Shen, Pavlo Molchanov, and Jose M. Alvarez. Adaptive sharpness-aware pruning for robust sparse networks. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=QFYVVwiAM8.

[2] Jose Blanchet, Peng Cui, Jiajin Li, and Jiashuo Liu. Stability evaluation through distributional perturbation analysis. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 4140–4159. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/blanchet24a.html.

[3] Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

[4] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. ARC-E (Easy) and ARC-C (Challenge) subsets.

[5] R.C. de Amorim et al. Normalization in classification: A comprehensive analysis and novel method. *arXiv preprint arXiv:2212.12343*, 2022. URL https://arxiv.org/abs/2212.12343.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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*, pages 4171–4186, Minneapolis, MN, USA, 2019. Association for Computational Linguistics.

[7] Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016.

[8] Ghadeer Jaradat, Mohammed Tolba, Ghada Alsuhli, Hani Saleh, Mahmoud Al-Qutayri, Thanos Stouraitis, and Baker Mohammad. Hybrid dynamic pruning: A pathway to efficient transformer inference. *arXiv preprint arXiv:2407.12893*, 2024. URL https://arxiv.org/abs/2407.12893.

[9] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[10] Pedro H Lima and Gerson Z Souza. A large comparison of normalization methods on time series. *ResearchGate Preprint*, 2023. URL https://www.researchgate.net/publication/373321082.

[11] Meta AI. meta-llama/llama-2-7b-hf. https://huggingface.co/meta-llama/Llama-2-7b-hf, 2023. Hugging Face model card; accessed 2025-07-28.

[12] Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[13] Saurav Muralidharan, Sharath Turuvekere Sreenivas, Raviraj Joshi, Marcin Chochowski, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, Jan Kautz, and Pavlo Molchanov. Compact language models via pruning and knowledge distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/4822991365c962105b1b95b1107d30e5-Paper-Conference.pdf.

[14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, 2017.

[16] 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)*, page to appear. Association for Computational Linguistics, 2019.

[17] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018. URL https://arxiv.org/abs/1804.07461.

[18] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019. Association for Computational Linguistics.

[19] Shuai Zhai, Jian Li, Zihang Wang, Yuhuai Wu, and Qiang Wang. Stabilizing transformer training by preventing attention entropy collapse. *arXiv preprint arXiv:2301.12916*, 2023.

Appendix A. Proofs and Details

Overview. This appendix collects the formal analyses that underpin Section 3.2 and fixes notation and technical conventions used throughout. We first develop loss-increase control for head pruning via gradient-based head importance (HIS), derive operator-norm curvature bounds for the cross-entropy objective, and justify the first-order approximation by quantifying when quadratic terms are negligible (Appendix A.1).

We then turn to generalization: starting from token-averaged notation and a neighboring-dataset construction, we link perturbations of attention distributions to output deviations and establish an entropy–total variation inequality that couples attention entropy with stability, culminating in a stability–generalization connection and practical constraints (Appendix A.2).

Building on these components, we present a risk upper bound whose surrogate minimization yields the proposed HIES objective and clarifies its role as a principled pruning criterion (Appendix A.3). Finally, we prove an orthogonality result between the centered HIS and entropy directions, showing their complementarity and explaining why combining the two signals improves robustness across pruning regimes (Appendix A.4).

Collectively, these results provide (i) tight loss-control guarantees under operator-norm curvature, (ii) an entropy-based route from stability to generalization, and (iii) a unified risk-motivated justification for HIES.

A.1. Loss-increase Control via Head Importance (Section 3.2.1)

A.1.1. LOSS BOUND WITH OPERATOR-NORM CONTROL

We write $\beta_y := \|\nabla^2_{\mathbf{y}}\mathcal{L}\|_2$ for the operator-norm curvature at the representation $\mathbf{y}$.
Consider the second-order Taylor expansion in $\mathbf{y}$:

$$\mathcal{L}(\mathbf{y}+\delta\mathbf{y}) \le \mathcal{L}(\mathbf{y}) + \langle \nabla_{\mathbf{y}}\mathcal{L},\, \delta\mathbf{y}\rangle_F + \frac{1}{2}\,\delta\mathbf{y}^\top \nabla^2_{\mathbf{y}}\mathcal{L}\,\delta\mathbf{y}.$$

Taking expectations and using the operator-norm bound yields

$$\Delta\mathcal{L} \;:=\; \mathbb{E}_{\mathcal{D}}\big[\mathcal{L}(\mathbf{y}+\delta\mathbf{y})-\mathcal{L}(\mathbf{y})\big] \;\le\; \mathbb{E}\big[\langle \nabla_{\mathbf{y}}\mathcal{L},\, \delta\mathbf{y}\rangle_F\big] + \frac{\beta_y}{2}\,\mathbb{E}\big[\|\delta\mathbf{y}\|_F^2\big].$$

Since $\delta\mathbf{y} = \mathrm{Concat}(\delta A_1,\ldots,\delta A_H)$, $\|\delta\mathbf{y}\|_F^2 = \sum_h \|\delta A_h\|_F^2$ and $\delta A_h = -(1-m_h)A_h$, the quadratic term equals $\frac{\beta_y}{2}\sum_h (1-m_h)\|A_h\|_F^2$. For the first-order term, using the absolute value in the HIS definition and head-wise triangle inequality,

$$\mathbb{E}\big[\langle \nabla_{\mathbf{y}}\mathcal{L},\, \delta\mathbf{y}\rangle_F\big] = -\sum_h (1-m_h)\,\mathbb{E}\big[\langle \nabla_{A_h}\mathcal{L},\, A_h\rangle_F\big] \;\le\; \sum_h (1-m_h)\,\mathrm{HIS}_h,$$

hence

$$\Delta\mathcal{L} \;\le\; \sum_{h=1}^{H}(1-m_h)\,\mathrm{HIS}_h \;+\; \frac{\beta_y}{2}\sum_{h=1}^{H}(1-m_h)\,\|A_h\|_F^2. \tag{6}$$

Plug-ins (default: binary).

Binary (Sigmoid) CE. $\beta_y \le \frac{1}{4}\|W^O\|_2^2,$

$$\Rightarrow \quad \Delta\mathcal{L} \le \sum_h (1-m_h)\,\mathrm{HIS}_h \;+\; \frac{1}{8}\,\|W^O\|_2^2 \sum_h (1-m_h)\,\|A_h\|_F^2.$$

Multiclass softmax CE. $\beta_y \le \frac{1}{2}\|W^O\|_2^2,$

$$\Rightarrow \quad \Delta\mathcal{L} \le \sum_h (1-m_h)\,\mathrm{HIS}_h \;+\; \frac{1}{4}\,\|W^O\|_2^2 \sum_h (1-m_h)\,\|A_h\|_F^2.$$

A.1.2. NORMS AND INNER PRODUCTS

We use token-averaged Frobenius norms and inner products: $\|A_h\|_F^2 = \frac{1}{n}\sum_{i=1}^n \|A_h(i)\|_2^2$ and $\langle U,V\rangle_F = \frac{1}{n}\sum_{i=1}^n \langle U(i),V(i)\rangle$ (*with batching*: replace $\frac{1}{n}$ by $\frac{1}{Bn}$).

A.1.3. ESTIMATING $\|W^O\|_2$ AND BLOCKWISE NORMS VIA POWER ITERATION

The spectral norm of a matrix $M \in \mathbb{R}^{m\times n}$ is defined as

$$\|M\|_2 := \max_{\|x\|_2=1} \|Mx\|_2,$$

which measures the maximum ℓ_2-amplification factor over all unit vectors. By the singular value decomposition (SVD), $M = U\Sigma V^\top$, where $\Sigma = \mathrm{diag}(\sigma_1,\sigma_2,\dots)$ with $\sigma_1 \ge \sigma_2 \ge \cdots \ge 0$, we have

$$\|M\|_2 = \sigma_{\max}(M),$$

i.e., the spectral norm equals the largest singular value. This follows since U and V are orthogonal and preserve the ℓ_2-norm, so the maximization reduces to aligning x with the right singular vector corresponding to $\sigma_{\max}$. Exact computation via full SVD costs $O(\min\{m^2n, mn^2\})$, which is prohibitive for large M. Instead, we approximate $\sigma_{\max}(M)$ using the *power iteration* method: starting from a random unit vector v_0, iterate

$$u_t \leftarrow \frac{Mv_t}{\|Mv_t\|_2}, \qquad v_{t+1} \leftarrow \frac{M^\top u_t}{\|M^\top u_t\|_2}.$$

After T iterations, $\|Mv_T\|_2$ converges to $\sigma_{\max}(M)$, and v_T approximates the corresponding right singular vector. We apply this procedure to W^O and, for a tighter quadratic term in our bound, to its head-wise column blocks $W^O_{(:,\mathcal{I}_h)}$, forming

$$\sum_h (1-m_h)\|W^O_{(:,\mathcal{I}_h)}\|_2^2\,\|A_h\|_F^2.$$

Here, $\mathcal{I}_h \subset \{1,\dots,d\}$ denotes the set of column indices in W^O corresponding to the d_v output dimensions of head h. Thus, $W^O_{(:,\mathcal{I}_h)} \in \mathbb{R}^{d\times d_v}$ is the column block of W^O mapping the d_v-dimensional output of head h to the d-dimensional model space.

A.1.4. CROSS-ENTROPY CURVATURE AND PROPAGATION TO $\mathbf{y}$

Logit-space Hessian (binary vs. multiclass). Binary (sigmoid) CE. For a single logit z with $p = \sigma(z)$,

$$\frac{d^2\mathcal{L}}{dz^2} \;=\; p(1-p) \;\le\; \tfrac{1}{4},$$

hence $\|\nabla^2_{\mathbf{z}}\mathcal{L}\|_2 \le \frac{1}{4}$.

Multiclass softmax CE. For logits $\mathbf{z} \in \mathbb{R}^C$ and $\mathbf{p} = \mathrm{softmax}(\mathbf{z})$,

$$\nabla^2_{\mathbf{z}}\mathcal{L} \;=\; \mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top, \quad \|\nabla^2_{\mathbf{z}}\mathcal{L}\|_2 \;\le\; \tfrac{1}{2}.$$

Proof sketch. For any unit vector v, $v^\top(\mathrm{diag}(p) - pp^\top)v = \sum_i p_i v_i^2 - (\sum_i p_i v_i)^2 = \mathrm{Var}_p(v)$. By Popoviciu's inequality, $\mathrm{Var}_p(v) \le \frac{(\max_i v_i - \min_i v_i)^2}{4} \le \frac{1}{2}$ for $\|v\|_2 = 1$. Tightness holds at $C = 2$, $p = (\frac{1}{2}, \frac{1}{2})$.

Mapping through W^O. With the immediate linear projection $\mathbf{z} = \mathbf{y}W^O$,

$$\nabla^2_{\mathbf{y}}\mathcal{L} \;=\; (W^O)^\top\,\nabla^2_{\mathbf{z}}\mathcal{L}\,W^O, \qquad \beta_y := \|\nabla^2_{\mathbf{y}}\mathcal{L}\|_2 \;\le\; \begin{cases} \frac{1}{4}\,\|W^O\|_2^2, & \text{binary CE,} \\ \frac{1}{2}\,\|W^O\|_2^2, & \text{multiclass softmax CE.} \end{cases}$$

(A blockwise refinement replaces $\|W^O\|_2$ by $\|W^O_{(:,\mathcal{I}_h)}\|_2$ head-wise.)

A.1.5. WHY THE QUADRATIC TERM IS TYPICALLY NEGLIGIBLE

We first note

$$\mathrm{HIS}_h \;=\; \mathbb{E}\big[\,|\langle\nabla_{A_h}\mathcal{L}, A_h\rangle_F|\,\big] \;=\; \mathbb{E}\big[\,|\cos\phi_h|\,\|\nabla_{A_h}\mathcal{L}\|_F\,\|A_h\|_F\,\big],$$

where the expectation is over $x \sim \mathcal{D}$ with token-averaging as in Appendix A.1.2. Assume there exists $g > 0$ such that, for all heads under consideration,

$$\mathrm{HIS}_h \;\ge\; g\,\mathbb{E}\big[\|A_h\|_F\big],$$

e.g., define

$$g \;:=\; \min_{h\in\{1,\dots,H\}}\; \mathbb{E}\big[\,|\cos\phi_h|\,\|\nabla_{A_h}\mathcal{L}\|_F\,\big],$$

where $\cos\phi_h \frac{\langle\nabla_{A_h}\mathcal{L}, A_h\rangle_F}{\|\nabla_{A_h}\mathcal{L}\|_F\,\|A_h\|_F}$ denotes the cosine alignment between the head's gradient and activation.

Then

$$\frac{\text{quadratic}}{\text{first-order}} \;\le\; \frac{\frac{\beta_y}{2}\sum_h(1-m_h)\,\mathbb{E}[\|A_h\|_F^2]}{\sum_h(1-m_h)\,\mathrm{HIS}_h} \;\le\; \frac{\beta_y}{2}\cdot\frac{\max_h\mathbb{E}[\|A_h\|_F]}{g},$$

where the second inequality uses $\sum_h(1-m_h)\mathbb{E}[\|A_h\|_F^2] \le \big(\max_h\mathbb{E}[\|A_h\|_F]\big)\sum_h(1-m_h)\mathbb{E}[\|A_h\|_F]$ and the per-head lower bound $\mathrm{HIS}_h \ge g\,\mathbb{E}[\|A_h\|_F]$.

Recalling $\beta_y \le c\,\|W^O\|_2^2$ with $c = \frac{1}{4}$ for binary CE and $c = \frac{1}{2}$ for multiclass softmax CE (cf. Appendix A.1.4), we obtain

$$\frac{\text{quadratic}}{\text{first-order}} \le \frac{c}{2}\,\|W^O\|_2^2 \cdot \frac{\max_h \mathbb{E}[\|A_h\|_F]}{g}.$$

A blockwise refinement further tightens this by replacing $\|W^O\|_2^2$ with $\max_h \|W^O_{(:,\mathcal{I}_h)}\|_2^2$. Since (i) LayerNorm controls token-wise activation scales (thus $\max_h \mathbb{E}[\|A_h\|_F]$), and (ii) g is bounded away from zero under non-degenerate alignment, the ratio is typically small. Hence the first-order term dominates in practice, while the second-order term remains explicitly controlled by the plug-in bounds in Appendix A.1.1.

A.1.6. REMARKS ON HIS WITH ABSOLUTE VALUES

The absolute value in (2) is part of the definition to prevent cancellation across samples; consequently, the triangle inequality turns the first-order term into an additive upper bound $\sum_h (1 - m_h)\,\mathrm{HIS}_h$ (cf. Appendix A.1.1). If $\langle \nabla_{A_h}\mathcal{L}, A_h\rangle_F < 0$ on some samples, masking that head could locally decrease the loss; the metric remains conservative by construction.

A.2. Generalization Gap and Attention Entropy (Section 3.2.2)

A.2.1. NOTATION AND TOKEN AVERAGING

For head h and query token $t \in \{1, \dots, n(x)\}$, let $\boldsymbol{\alpha}_t^{(h)}(x) \in \Delta^{n(x)-1}$ denote the attention distribution over keys, and $H(\mathbf{p}) := -\sum_j p_j \log p_j$ the entropy. Define the *token-averaged, length-normalized entropy* and *deficit* by

$$\mathrm{AE}_h(x) := \frac{1}{n(x)\,\log n(x)} \sum_{t=1}^{n(x)} H\big(\boldsymbol{\alpha}_t^{(h)}(x)\big) \in [0,1],$$

$$\mathrm{AD}_h(x) := \frac{1}{n(x)\,\log n(x)} \sum_{t=1}^{n(x)} \Big(\log n(x) - H\big(\boldsymbol{\alpha}_t^{(h)}(x)\big)\Big) \;=\; 1 - \mathrm{AE}_h(x) \in [0,1].$$

For neighboring datasets $(\mathcal{S}, \mathcal{S}')$, write the symmetric aggregation

$$\overline{\mathrm{AD}}_h(x) \;:=\; \tfrac{1}{2}\big(\mathrm{AD}_h(x) + \mathrm{AD}'_h(x)\big).$$

All token averages exclude padding positions and use the effective context length for causal masking(cf. Appendix A.2.1). Here $(\mathcal{S}, \mathcal{S}')$ are *neighboring* datasets that differ in one example.

A.2.2. NEIGHBORING DATASETS AND WHY THEY APPEAR

We call two datasets $S = (z_1, \dots, z_N)$ and $S' = (z_1, \dots, z_{i-1}, z'_i, z_{i+1}, \dots, z_N)$ *neighboring* if they differ in exactly one example.

Why neighboring datasets?

- **Symmetrization.** Introduce an i.i.d. ghost sample $S' \sim \mathcal{D}^N$ to rewrite the expected generalization gap as an average of sample-wise differences, e.g., $\mathbb{E}_{S,S'}\big[\frac{1}{N}\sum_{i=1}^N\big(\ell(f_S; z_i) - \ell(f_{S'}; z'_i)\big)\big]$, which is amenable to concentration and stability arguments.
- **Replace-one stability.** Measure sensitivity to a single replacement by comparing f_S with $f_{S^{(i\leftarrow z')}}$, where $S^{(i\leftarrow z')}$ replaces z_i by z'_i; under γ-uniform stability and bounded loss B, this yields $\mathbb{E}_S[\mathcal{G}(S)] \le 2\gamma + \frac{B}{N}$.
- **Symmetric inequalities.** Our entropy–total variation (TV) control is symmetric in two distributions (α, α'); we thus aggregate via $\overline{\mathrm{AD}}_h(x) := \frac{1}{2}\big(\mathrm{AD}_h(x) + \mathrm{AD}'_h(x)\big)$, which streamlines notation and tightens constants in the perturbation bound.

A.2.3. FROM ATTENTION PERTURBATION TO OUTPUT PERTURBATION

For token t, the head output is $a_h(t) = \big(\boldsymbol{\alpha}_t^{(h)}\big)^\top V_h \in \mathbb{R}^{d_v}$, hence for neighboring datasets,

$$\Delta_h(t) \;:=\; a_h(t) - a'_h(t) \;=\; \big(\boldsymbol{\alpha}_t^{(h)} - \boldsymbol{\alpha}_t'^{(h)}\big)^\top V_h.$$

With $\|V_h\|_{\infty\to 2} := \max_j \|V_h(j,:)\|_2$ and $\|V_h\|_{\infty\to 2} \le M$,

$$\|\Delta_h(t)\|_2 \;\le\; \|\boldsymbol{\alpha}_t^{(h)} - \boldsymbol{\alpha}_t'^{(h)}\|_1 \cdot \|V_h\|_{\infty\to 2} \;\le\; M\,\|\boldsymbol{\alpha}_t^{(h)} - \boldsymbol{\alpha}_t'^{(h)}\|_1. \tag{7}$$

Averaging over tokens and applying the mask m_h,

$$\|\Delta(x)\|_2 \;:=\; \frac{1}{n(x)} \sum_{t=1}^{n(x)} \sum_{h=1}^{H} (1 - m_h)\,\|\Delta_h(t)\|_2.$$

A.2.4. ENTROPY–TOTAL VARIATION (TV) CONTROL

Lemma 4 (Entropy–TV inequality) *For* $\mathbf{p}, \mathbf{q} \in \Delta^{n-1}$ *and* $\mathbf{u}$ *uniform,* $\|\mathbf{p} - \mathbf{q}\|_1^2 \leq 4\big[H(\mathbf{u}) - H(\mathbf{p}) + H(\mathbf{u}) - H(\mathbf{q})\big]$.

Proof Triangle inequality and $(a+b)^2 \leq 2(a^2+b^2)$ give $\|\mathbf{p} - \mathbf{q}\|_1^2 \leq 2(\|\mathbf{p} - \mathbf{u}\|_1^2 + \|\mathbf{q} - \mathbf{u}\|_1^2)$. Pinsker w.r.t. $\mathbf{u}$ yields $\|\mathbf{p} - \mathbf{u}\|_1^2 \leq 2(\log n - H(\mathbf{p}))$ and likewise for $\mathbf{q}$. ■

Applying Lemma 4 to (7) token-wise and averaging,

$$\frac{1}{n(x)} \sum_{t=1}^{n(x)} \|\boldsymbol{\alpha}_t^{(h)} - \boldsymbol{\alpha}_t'^{(h)}\|_1 \;\leq\; \sqrt{\frac{1}{n(x)} \sum_{t=1}^{n(x)} \|\boldsymbol{\alpha}_t^{(h)} - \boldsymbol{\alpha}_t'^{(h)}\|_1^2} \;\leq\; \sqrt{8 \log n(x)}\; \sqrt{\overline{\mathrm{AD}}_h(x)}.$$

Therefore,

$$\|\Delta(x)\|_2 \;\leq\; M\sqrt{8\log n(x)} \sum_{h=1}^{H} (1 - m_h)\, \sqrt{\overline{\mathrm{AD}}_h(x)}.$$

By Cauchy–Schwarz and $\sum_h (1 - m_h) = H\rho$,

$$\|\Delta(x)\|_2 \;\leq\; \underbrace{\sqrt{8}\, M\, \sqrt{H\rho\, \log n(x)}}_{=:\, C_{\mathrm{AE}}(x)} \cdot \sqrt{\sum_{h=1}^{H} (1 - m_h)\, \overline{\mathrm{AD}}_h(x)}. \tag{8}$$

A.2.5. STABILITY AND GENERALIZATION

Let $\gamma := L_\ell\, \mathbb{E}_{\mathcal{S}}[\|\Delta(x)\|_2]$. By on-average replace-one stability [3, Def. 6],

$$\mathbb{E}_{\mathcal{S}}\big[\mathcal{G}(\mathcal{S})\big] \leq 2\gamma, \qquad \mathcal{G}(\mathcal{S}) \leq 2\gamma + \tfrac{B}{N}.$$

Using (8) and Jensen for $\sqrt{\cdot}$,

$$\gamma \;\leq\; L_\ell\, \mathbb{E}_{\mathcal{S}}\big[C_{\mathrm{AE}}(x)\big] \cdot \sqrt{\sum_{h=1}^{H} (1 - m_h)\, \mathbb{E}_{\mathcal{S}}\big[\overline{\mathrm{AD}}_h(x)\big]}.$$

Taking a representative n (e.g., average/max effective length) yields the main-text constant $C_{\mathrm{AE}} = \sqrt{8}\, M\, \sqrt{H\rho\, \log n}$ and Eq. (5).

A.2.6. CONSTANTS AND PRACTICAL REMARKS

- **Operator norm.** $\|V_h\|_{\infty\to 2} := \max_j \|V_h(j,:)\|_2$; take $M := \max_h \|V_h\|_{\infty \to 2}$ (controlled by LayerNorm/weight norms).
- **Sequence length.** For padding/causal masking, replace $n(x)$ by the effective context length; averages exclude padded positions.
- **Deficit aggregation.** On-average: $\overline{\mathrm{AD}}_h = \frac{1}{2}(\mathrm{AD}_h + \mathrm{AD}'_h)$; Uniform: $\overline{\mathrm{AD}}_h = \max\{\mathrm{AD}_h, \mathrm{AD}'_h\}$.
- **Do not pool entropies.** Since using $H(\frac{1}{n}\sum_t \alpha_t)$ can underestimate deficit (Jensen) and weaken control, token-wise entropies are required.

A.3. Risk Upper Bound and HIES Minimization (Section 3.2.3)

Proof Let $\mathrm{supp}(m) := \{h : m_h = 1\}$ denote the set of retained heads. Suppose an admissible mask m' with $|\,\mathrm{supp}(m')| = k$ is not optimal. Then there exist $i \in \mathrm{supp}(m')$ and $j \notin \mathrm{supp}(m')$ such that HIES$_j$ $>$ HIES$_i$. Consider the mask $\tilde{m}$ that swaps i and j (retain j, prune i); the constraint in (2) is preserved. The objective in (1) changes by

$$\Delta\mathcal{R} = \big[\text{HIES}_i\big] - \big[\text{HIES}_j\big] < 0,$$

since j was contributing to the sum (pruned) and i was not (retained). Hence $\tilde{m}$ has a strictly smaller objective, contradicting the minimality of m'. Therefore retaining the k heads with the largest HIES is optimal; equivalently, pruning the $H - k$ smallest HIES is optimal. ■

A.4. Orthogonality and Complementarity (Section 3.2.4)

Preliminaries. For head h, let $\boldsymbol{\alpha}^{(h)} \in \Delta^{n-1}$ be the attention probability vector, $V_h \in \mathbb{R}^{n \times d_v}$ the value matrix, and

$$A_h \;=\; \boldsymbol{\alpha}^{(h)} V_h \;\in\; \mathbb{R}^{1 \times d_v}.$$

Define

$$g_h \;:=\; V_h \big(\nabla_{A_h} \mathcal{L}\big)^\top \;\in\; \mathbb{R}^n, \qquad \mathrm{HIS}_h \;=\; \big|\boldsymbol{\alpha}^{(h)\top} g_h\big|, \qquad \mathrm{AE}_h \;=\; -\sum_{j=1}^{n} \alpha_j^{(h)} \log \alpha_j^{(h)}.$$

Gradients w.r.t. attention (interior points). For $\alpha_j^{(h)} > 0$,

$$\nabla_{\boldsymbol{\alpha}^{(h)}} \mathrm{HIS}_h = \mathrm{sign}\big(\boldsymbol{\alpha}^{(h)\top} g_h\big)\, g_h, \qquad \nabla_{\boldsymbol{\alpha}^{(h)}} \mathrm{AE}_h = -\big(\mathbf{1} + \log \boldsymbol{\alpha}^{(h)}\big),$$

where $\log$ is applied elementwise. (At $\boldsymbol{\alpha}^{(h)\top} g_h = 0$, any subgradient in $\{s\, g_h : s \in [-1, 1]\}$ is valid; this does not affect the result in expectation.)

Simplex projection. Since $\boldsymbol{\alpha}^{(h)} \in \Delta^{n-1}$, we project onto the tangent space with $P := I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ and define

$$\widetilde{\nabla}\mathrm{HIS}_h := P\, \nabla \mathrm{HIS}_h, \qquad \widetilde{\nabla}\mathrm{AE}_h := P\, \nabla \mathrm{AE}_h.$$

Proof By definition, $u_h := \mathrm{sign}(\boldsymbol{\alpha}^{(h)\top} g_h)\, g_h$, $v_h := \mathbf{1} + \log \boldsymbol{\alpha}^{(h)}$, and $\tilde{u}_h := P u_h$, $\tilde{v}_h := P v_h$. Then

$$\widetilde{\nabla}_{\boldsymbol{\alpha}^{(h)}} \mathrm{HIS}_h = P\, \nabla_{\boldsymbol{\alpha}^{(h)}} \mathrm{HIS}_h = \tilde{u}_h, \qquad \widetilde{\nabla}_{\boldsymbol{\alpha}^{(h)}} \mathrm{AE}_h = P\, \nabla_{\boldsymbol{\alpha}^{(h)}} \mathrm{AE}_h = -\tilde{v}_h.$$

Hence

$$\mathbb{E}_{\mathcal{S}}\big[\, \langle \widetilde{\nabla}\mathrm{HIS}_h,\, \widetilde{\nabla}\mathrm{AE}_h \rangle \,\big] = \mathbb{E}_{\mathcal{S}}\big[\langle \tilde{u}_h,\, -\tilde{v}_h \rangle\big] = -\,\mathrm{tr}\Big(\, \mathbb{E}_{\mathcal{S}}\big[\tilde{u}_h \tilde{v}_h^\top\big]\Big).$$

Decomposing the second moment,

$$\mathbb{E}_{\mathcal{S}}\big[\tilde{u}_h \tilde{v}_h^\top\big] \;=\; \mathrm{Cov}(\tilde{u}_h, \tilde{v}_h) + \mathbb{E}_{\mathcal{S}}[\tilde{u}_h]\, \mathbb{E}_{\mathcal{S}}[\tilde{v}_h]^\top.$$

Under $\mathrm{Cov}(\tilde{u}_h, \tilde{v}_h) = 0$ and $\langle \mathbb{E}_{\mathcal{S}}[\tilde{u}_h],\, \mathbb{E}_{\mathcal{S}}[\tilde{v}_h] \rangle = 0$ (*or* the stronger $\mathbb{E}_{\mathcal{S}}[\tilde{u}_h] = 0$), we obtain

$$\mathbb{E}_{\mathcal{S}}\big[\, \langle \widetilde{\nabla}\mathrm{HIS}_h,\, \widetilde{\nabla}\mathrm{AE}_h \rangle \,\big] \;=\; 0.$$

■

Technical remarks. (i) At $\boldsymbol{\alpha}^{(h)\top} g_h = 0$, use any subgradient of $|\cdot|$ for $\nabla_{\boldsymbol{\alpha}^{(h)}} \mathrm{HIS}_h$. (ii) Since $\boldsymbol{\alpha}^{(h)} = \mathrm{softmax}(\cdot)$, we have $\alpha_j^{(h)} > 0$, so $\log \boldsymbol{\alpha}^{(h)}$ (elementwise) is well-defined. (iii) "$\mathrm{Cov}(x, y) = 0$" denotes the *cross-covariance matrix* being zero, not merely componentwise uncorrelatedness. (iv) If one omits the projection P, the same argument applies with u_h, v_h replacing $\tilde{u}_h, \tilde{v}_h$ under the analogous conditions $\mathrm{Cov}(u_h, v_h) = 0$ and $\langle \mathbb{E}_{\mathcal{S}}[u_h],\, \mathbb{E}_{\mathcal{S}}[v_h] \rangle = 0$.

Appendix B. Experimental Setup

Model. We use publicly available BERT checkpoints that have been fine-tuned and released by prior work [6], and Meta's LLaMA 2–7B–hf checkpoint from Hugging Face [11].

Datasets. We evaluate on two standard benchmarks: **GLUE** [17] and **HellaSwag** [18] and the AI2 Reasoning Challenge—**ARC-e/ARC-c** [4].

Baselines.

- **Random**: prune attention heads uniformly at random.
- **HIS**: prune heads with smaller head-importance first [12, 16]. We use

$$\mathrm{HIS}_h \;=\; \mathbb{E}_{x\sim\mathcal{D}}\left|\frac{\partial\mathcal{L}(x)}{\partial\xi_h}\right| \;=\; \mathbb{E}_{x\sim\mathcal{D}}\left|\,\big\langle\nabla_{A_h(x)}\mathcal{L}(x),\, A_h(x)\big\rangle_F\,\right|.$$

- **Attention Deficit (AD;** $1-$**Attention Entropy)**: prune heads with smaller attention entropy first (i.e., more concentrated attention patterns). Writing the attention-entropy (AE) of head h as

$$\mathrm{AE}_h \;=\; H\big(\boldsymbol{\alpha}^{(h)}\big) \;=\; -\sum_{i=1}^{n}\alpha_i^{(h)}\,\log\alpha_i^{(h)},$$

  we use the normalized deficit

$$\mathrm{AD}_h \;=\; 1 \;-\; \frac{H(\boldsymbol{\alpha}^{(h)})}{\log n},$$

  so that lower entropy corresponds to larger AD_h; pruning proceeds from lower entropy (equivalently larger AD_h).

Appendix C. Additional Experimental Results

C.1. Orthogonality Analysis

We provide an empirical sanity check supporting the assumptions above. Experiments use TinyBERT on SST-2 (`Vishnou/TinyBERT_SST2`). For each head we compute layerwise-normalized HIS and attention-entropy (AE) scores, stack them into vectors u and v, and form the centered versions $\tilde{u}$ and $\tilde{v}$ by subtracting each vector's mean (a finite-sample proxy for projection onto the zero-sum subspace). The following sample statistics were obtained:

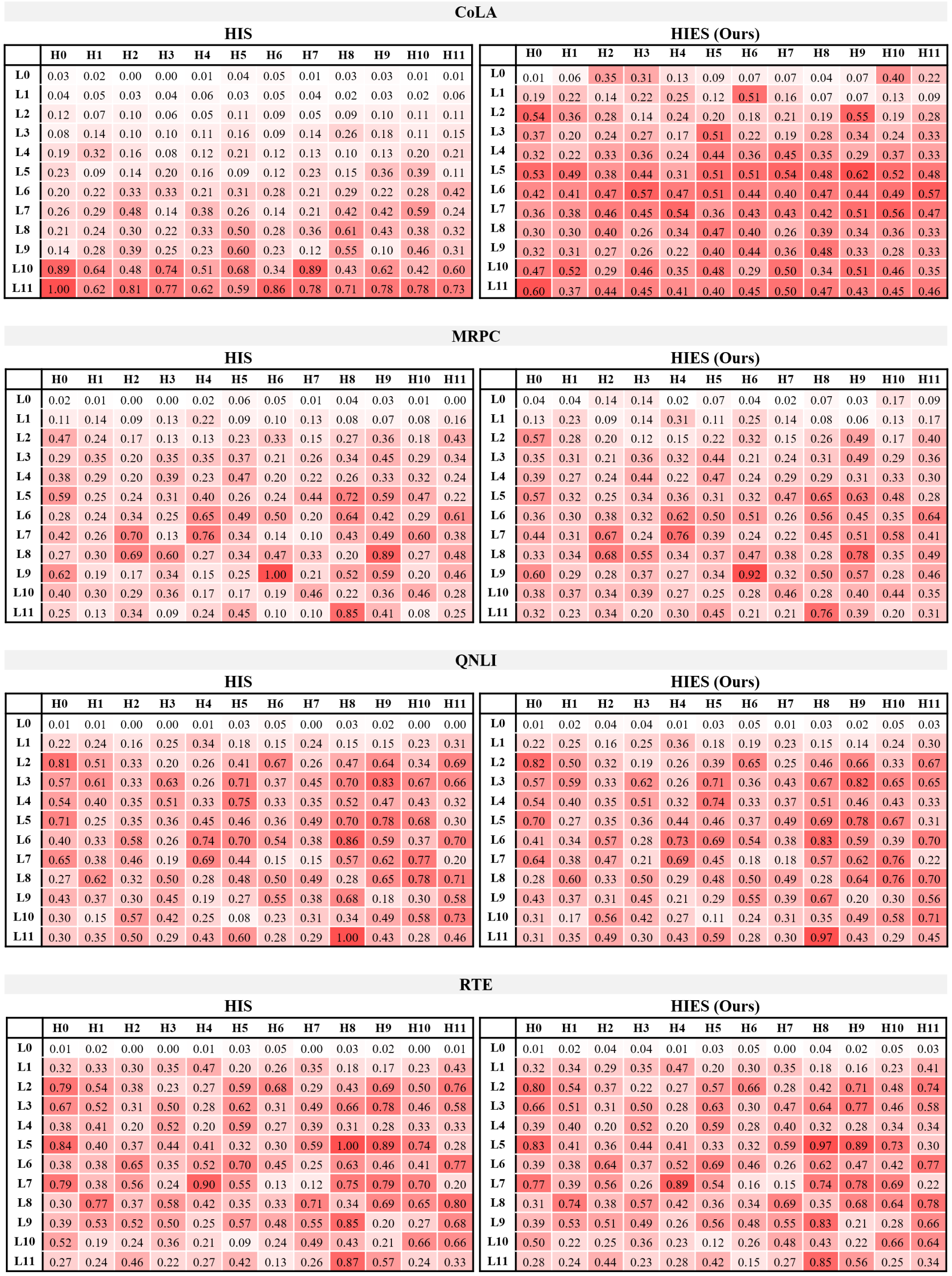
$$\widehat{\mathrm{Cov}}(\tilde{u}, \tilde{v}) = 0.030853, \qquad \overline{u} = 3.73 \times 10^{-9}, \qquad \overline{v} = 1.61 \times 10^{-8}.$$

Consequently,

$$\widehat{\mathbb{E}}\big[\langle \tilde{u},\, -\tilde{v} \rangle\big] \;=\; -0.030853 \;=\; -\big(\widehat{\mathrm{Cov}}(\tilde{u}, \tilde{v}) + \overline{u}\,\overline{v}\big) \quad \text{(up to numerical precision)}.$$

The covariance magnitude is small on this batch, indicating weak coupling between the two directions and lending empirical support to the "***(near) uncorrelatedness***" assumption. We recommend reporting the same diagnostics averaged over multiple batches to reduce sampling noise and to provide confidence intervals.

C.2. Heatmap of Importance Scores and Pruning Results



CoLA

HIS

	H0	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11
L0	0.03	0.02	0.00	0.00	0.01	0.04	0.05	0.01	0.03	0.03	0.01	0.01
L1	0.04	0.05	0.03	0.04	0.06	0.03	0.05	0.04	0.02	0.03	0.02	0.06
L2	0.12	0.07	0.10	0.06	0.05	0.11	0.09	0.05	0.09	0.10	0.11	0.11
L3	0.08	0.14	0.10	0.10	0.11	0.16	0.09	0.14	0.26	0.18	0.11	0.15
L4	0.19	0.32	0.16	0.08	0.12	0.21	0.12	0.13	0.10	0.13	0.20	0.21
L5	0.23	0.09	0.14	0.20	0.16	0.09	0.12	0.23	0.15	0.36	0.39	0.11
L6	0.20	0.22	0.33	0.33	0.21	0.31	0.28	0.21	0.29	0.22	0.28	0.42
L7	0.26	0.29	0.48	0.14	0.38	0.26	0.14	0.21	0.42	0.42	0.59	0.24
L8	0.21	0.24	0.30	0.22	0.33	0.50	0.28	0.36	0.61	0.43	0.38	0.32
L9	0.14	0.28	0.39	0.25	0.23	0.60	0.23	0.12	0.55	0.10	0.46	0.31
L10	0.89	0.64	0.48	0.74	0.51	0.68	0.34	0.89	0.43	0.62	0.42	0.60
L11	1.00	0.62	0.81	0.77	0.62	0.59	0.86	0.78	0.71	0.78	0.78	0.73

HIES (Ours)

	H0	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11
L0	0.01	0.06	0.35	0.31	0.13	0.09	0.07	0.07	0.04	0.07	0.40	0.22
L1	0.19	0.22	0.14	0.22	0.25	0.12	0.51	0.16	0.07	0.07	0.13	0.09
L2	0.54	0.36	0.28	0.14	0.24	0.20	0.18	0.21	0.19	0.55	0.19	0.28
L3	0.37	0.20	0.24	0.27	0.17	0.51	0.22	0.19	0.28	0.34	0.24	0.33
L4	0.32	0.22	0.33	0.36	0.24	0.44	0.36	0.45	0.35	0.29	0.37	0.33
L5	0.53	0.49	0.38	0.44	0.31	0.51	0.51	0.54	0.48	0.62	0.52	0.48
L6	0.42	0.41	0.47	0.57	0.47	0.51	0.44	0.40	0.47	0.44	0.49	0.57
L7	0.36	0.38	0.46	0.45	0.54	0.36	0.43	0.43	0.42	0.51	0.56	0.47
L8	0.30	0.30	0.40	0.26	0.34	0.47	0.40	0.26	0.39	0.34	0.36	0.33
L9	0.32	0.31	0.27	0.26	0.22	0.40	0.44	0.36	0.48	0.33	0.28	0.33
L10	0.47	0.52	0.29	0.46	0.35	0.48	0.29	0.50	0.34	0.51	0.46	0.35
L11	0.60	0.37	0.44	0.45	0.41	0.40	0.45	0.50	0.47	0.43	0.45	0.46

MRPC

HIS

	H0	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11
L0	0.02	0.01	0.00	0.00	0.02	0.06	0.05	0.01	0.04	0.03	0.01	0.00
L1	0.11	0.14	0.09	0.13	0.22	0.09	0.10	0.13	0.08	0.07	0.08	0.16
L2	0.47	0.24	0.17	0.13	0.13	0.23	0.33	0.15	0.27	0.36	0.18	0.43
L3	0.29	0.35	0.20	0.35	0.35	0.37	0.21	0.26	0.34	0.45	0.29	0.34
L4	0.38	0.29	0.20	0.39	0.23	0.47	0.20	0.22	0.26	0.33	0.32	0.24
L5	0.59	0.25	0.24	0.31	0.40	0.26	0.24	0.44	0.72	0.59	0.47	0.22
L6	0.28	0.24	0.34	0.25	0.65	0.49	0.50	0.20	0.64	0.42	0.29	0.61
L7	0.42	0.26	0.70	0.13	0.76	0.34	0.14	0.10	0.43	0.49	0.60	0.38
L8	0.27	0.30	0.69	0.60	0.27	0.34	0.47	0.33	0.20	0.89	0.27	0.48
L9	0.62	0.19	0.17	0.34	0.15	0.25	1.00	0.21	0.52	0.59	0.20	0.46
L10	0.40	0.30	0.29	0.36	0.17	0.17	0.19	0.46	0.22	0.36	0.46	0.28
L11	0.25	0.13	0.34	0.09	0.24	0.45	0.10	0.10	0.85	0.41	0.08	0.25

HIES (Ours)

	H0	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11
L0	0.04	0.04	0.14	0.14	0.02	0.07	0.04	0.02	0.07	0.03	0.17	0.09
L1	0.13	0.23	0.09	0.14	0.31	0.11	0.25	0.14	0.08	0.06	0.13	0.17
L2	0.57	0.28	0.20	0.12	0.15	0.22	0.32	0.15	0.26	0.49	0.17	0.40
L3	0.35	0.31	0.21	0.36	0.32	0.44	0.21	0.24	0.31	0.49	0.29	0.36
L4	0.39	0.27	0.24	0.44	0.22	0.47	0.24	0.29	0.29	0.31	0.33	0.30
L5	0.57	0.32	0.25	0.34	0.36	0.31	0.32	0.47	0.65	0.63	0.48	0.28
L6	0.36	0.30	0.38	0.32	0.62	0.50	0.51	0.26	0.56	0.45	0.35	0.64
L7	0.44	0.31	0.67	0.24	0.76	0.39	0.24	0.22	0.45	0.51	0.58	0.41
L8	0.33	0.34	0.68	0.55	0.34	0.37	0.47	0.38	0.28	0.78	0.35	0.49
L9	0.60	0.29	0.28	0.37	0.27	0.34	0.92	0.32	0.50	0.57	0.28	0.46
L10	0.38	0.37	0.34	0.39	0.27	0.25	0.28	0.46	0.28	0.40	0.44	0.35
L11	0.32	0.23	0.34	0.20	0.30	0.45	0.21	0.21	0.76	0.39	0.20	0.31

QNLI

HIS

	H0	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11
L0	0.01	0.01	0.00	0.00	0.01	0.03	0.05	0.00	0.03	0.02	0.00	0.00
L1	0.22	0.24	0.16	0.25	0.34	0.18	0.15	0.24	0.15	0.15	0.23	0.31
L2	0.81	0.51	0.33	0.20	0.26	0.41	0.67	0.26	0.47	0.64	0.34	0.69
L3	0.57	0.61	0.33	0.63	0.26	0.71	0.37	0.45	0.70	0.83	0.67	0.66
L4	0.54	0.40	0.35	0.51	0.33	0.75	0.33	0.35	0.52	0.47	0.43	0.32
L5	0.71	0.25	0.35	0.36	0.45	0.46	0.36	0.49	0.70	0.78	0.68	0.30
L6	0.40	0.33	0.58	0.26	0.74	0.70	0.54	0.38	0.86	0.59	0.37	0.70
L7	0.65	0.38	0.46	0.19	0.69	0.44	0.15	0.15	0.57	0.62	0.77	0.20
L8	0.27	0.62	0.32	0.50	0.28	0.48	0.50	0.49	0.28	0.65	0.78	0.71
L9	0.43	0.37	0.30	0.45	0.19	0.27	0.55	0.38	0.68	0.18	0.30	0.58
L10	0.30	0.15	0.57	0.42	0.25	0.08	0.23	0.31	0.34	0.49	0.58	0.73
L11	0.30	0.35	0.50	0.29	0.43	0.60	0.28	0.29	1.00	0.43	0.28	0.46

HIES (Ours)

	H0	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11
L0	0.01	0.02	0.04	0.04	0.01	0.03	0.05	0.01	0.03	0.02	0.05	0.03
L1	0.22	0.25	0.16	0.25	0.36	0.18	0.19	0.23	0.15	0.14	0.24	0.30
L2	0.82	0.50	0.32	0.19	0.26	0.39	0.65	0.25	0.46	0.66	0.33	0.67
L3	0.57	0.59	0.33	0.62	0.26	0.71	0.36	0.43	0.67	0.82	0.65	0.65
L4	0.54	0.40	0.35	0.51	0.32	0.74	0.33	0.37	0.51	0.46	0.43	0.33
L5	0.70	0.27	0.35	0.36	0.44	0.46	0.37	0.49	0.69	0.78	0.67	0.31
L6	0.41	0.34	0.57	0.28	0.73	0.69	0.54	0.38	0.83	0.59	0.39	0.70
L7	0.64	0.38	0.47	0.21	0.69	0.45	0.18	0.18	0.57	0.62	0.76	0.22
L8	0.28	0.60	0.33	0.50	0.29	0.48	0.50	0.49	0.28	0.64	0.76	0.70
L9	0.43	0.37	0.31	0.45	0.21	0.29	0.55	0.39	0.67	0.20	0.30	0.56
L10	0.31	0.17	0.56	0.42	0.27	0.11	0.24	0.31	0.35	0.49	0.58	0.71
L11	0.31	0.35	0.49	0.30	0.43	0.59	0.28	0.30	0.97	0.43	0.29	0.45

RTE

HIS

	H0	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11
L0	0.01	0.02	0.00	0.00	0.01	0.03	0.05	0.00	0.03	0.02	0.00	0.01
L1	0.32	0.33	0.30	0.35	0.47	0.20	0.26	0.35	0.18	0.17	0.23	0.43
L2	0.79	0.54	0.38	0.23	0.27	0.59	0.68	0.29	0.43	0.69	0.50	0.76
L3	0.67	0.52	0.31	0.50	0.28	0.62	0.31	0.49	0.66	0.78	0.46	0.58
L4	0.38	0.41	0.20	0.52	0.20	0.59	0.27	0.39	0.31	0.28	0.33	0.33
L5	0.84	0.40	0.37	0.44	0.41	0.32	0.30	0.59	1.00	0.89	0.74	0.28
L6	0.38	0.38	0.65	0.35	0.52	0.70	0.45	0.25	0.63	0.46	0.41	0.77
L7	0.79	0.38	0.56	0.24	0.90	0.55	0.13	0.12	0.75	0.79	0.70	0.20
L8	0.30	0.77	0.37	0.58	0.42	0.35	0.33	0.71	0.34	0.69	0.65	0.80
L9	0.39	0.53	0.52	0.50	0.25	0.57	0.48	0.55	0.85	0.20	0.27	0.68
L10	0.52	0.19	0.24	0.36	0.21	0.09	0.24	0.49	0.43	0.21	0.66	0.66
L11	0.27	0.24	0.46	0.22	0.27	0.42	0.13	0.26	0.87	0.57	0.24	0.33

HIES (Ours)

	H0	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11
L0	0.01	0.02	0.04	0.04	0.01	0.03	0.05	0.00	0.04	0.02	0.05	0.03
L1	0.32	0.34	0.29	0.35	0.47	0.20	0.30	0.35	0.18	0.16	0.23	0.41
L2	0.80	0.54	0.37	0.22	0.27	0.57	0.66	0.28	0.42	0.71	0.48	0.74
L3	0.66	0.51	0.31	0.50	0.28	0.63	0.30	0.47	0.64	0.77	0.46	0.58
L4	0.39	0.40	0.20	0.52	0.20	0.59	0.28	0.40	0.32	0.28	0.34	0.34
L5	0.83	0.41	0.36	0.44	0.41	0.33	0.32	0.59	0.97	0.89	0.73	0.30
L6	0.39	0.38	0.64	0.37	0.52	0.69	0.46	0.26	0.62	0.47	0.42	0.77
L7	0.77	0.39	0.56	0.26	0.89	0.54	0.16	0.15	0.74	0.78	0.69	0.22
L8	0.31	0.74	0.38	0.57	0.42	0.36	0.34	0.69	0.35	0.68	0.64	0.78
L9	0.39	0.53	0.51	0.49	0.26	0.56	0.48	0.55	0.83	0.21	0.28	0.66
L10	0.50	0.22	0.25	0.36	0.23	0.12	0.26	0.48	0.43	0.22	0.66	0.64
L11	0.28	0.24	0.44	0.23	0.28	0.42	0.15	0.27	0.85	0.56	0.25	0.34

Figure 1: Heatmaps of head-importance scores across four GLUE tasks (CoLA, MRPC, QNLI, RTE). Left: HIS; Right: HIES (ours). Rows = layers (L0–L11); columns = heads (H0–H11).

We analyze the pruning patterns and performance dynamics of HIS- and HIES-based methods across varying sparsity levels. This section highlights the fundamental distinctions in head selection strategies and the underlying mechanisms responsible for the observed performance inversion.

C.2.1. DIFFERENCE IN PRUNING PATTERNS

Pruning heatmaps (Fig. 2) reveal systematic differences between the methods. HIS-based pruning tends to remove heads primarily from the lower layers, producing an approximately bottom-up pattern consistent with its one-step gradient saliency. In contrast, HIES yields a more dispersed selection spanning lower, middle, and upper layers. We attribute this to the entropy-aware term, which leverages structural properties of the attention distribution (concentration vs. dispersion) in addition to gradient sensitivity, thereby promoting diversity across layers in pruning decisions.

C.2.2. PERFORMANCE INVERSION ACROSS SPARSITY REGIMES

We identify two distinct pruning regimes:

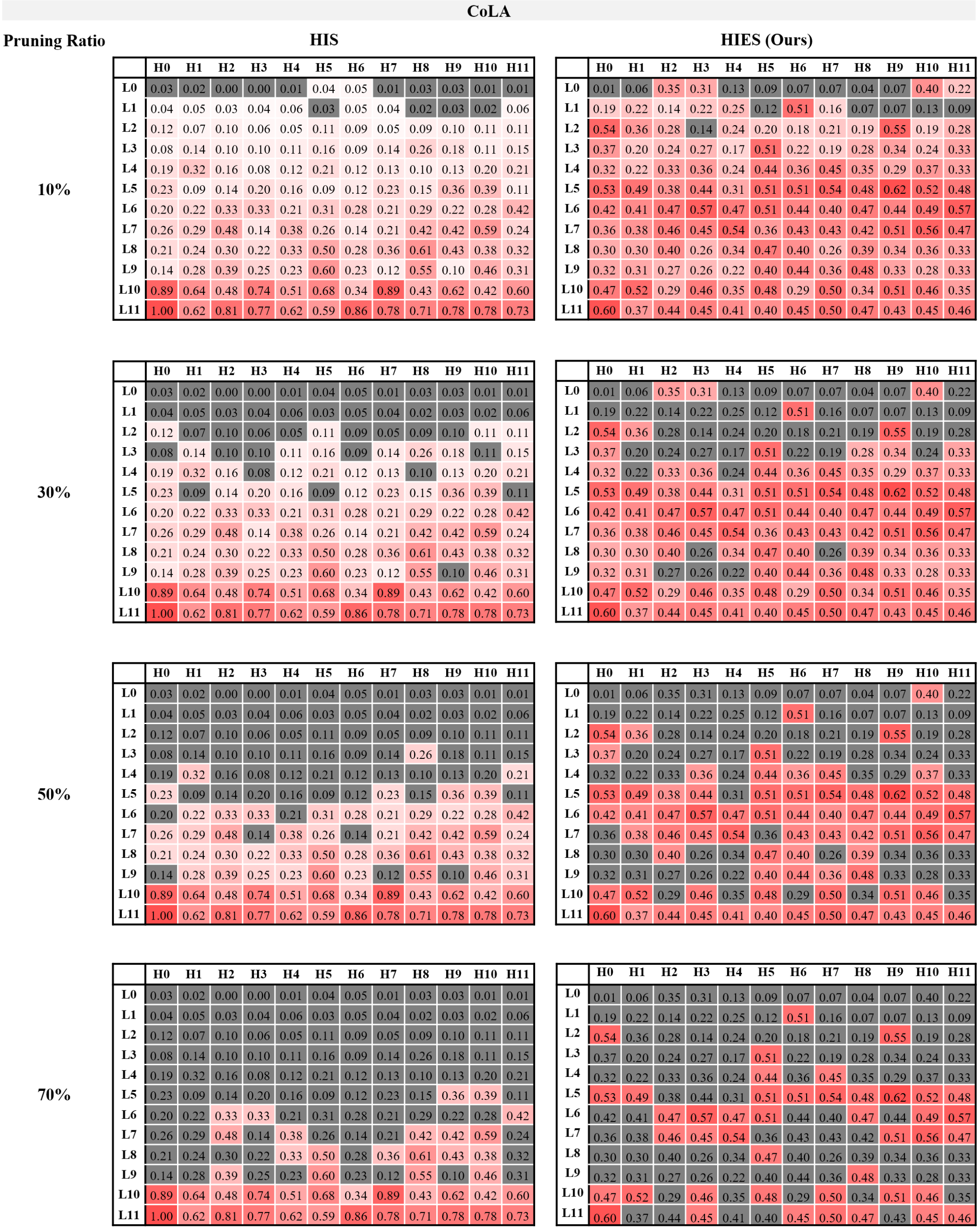
Redundancy Regime ($\leq 10\%$ pruning). In the early pruning phase, the model contains a substantial number of redundant heads. Here, gradient-based importance scores (HIS) are sufficient to identify and remove low-sensitivity heads, as they reflect the immediate (one-step) loss change. Consequently, HIS performs slightly better than HIES in both accuracy and stability under light pruning.

Specialization Regime ($\geq 30\%$ pruning). As pruning becomes more aggressive, redundant heads are mostly exhausted, and specialized heads begin to be targeted. In this regime, HIS alone struggles to distinguish critical heads from less important ones, as gradient magnitudes no longer capture long-term utility. In contrast, HIES leverages attention entropy to preferentially preserve highly concentrated (low-entropy) heads—which are typically more specialized—and prune high-entropy, less task-specific heads. This leads to superior accuracy and stability under higher pruning ratios.

Summary

- **Pruning $\leq 10\%$:** Redundancy regime ⇒ HIS outperforms HIES.
- **Pruning $\geq 30\%$:** Specialization regime ⇒ HIES outperforms HIS.

These findings demonstrate that HIS and HIES prioritize head preservation differently—HIS reflects short-horizon gradient sensitivity, whereas HIES incorporates extended inference-time stability by preserving low-entropy specialized heads.



CoLA

Pruning Ratio | **HIS** | **HIES (Ours)**

10%

HIS

	H0	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11
L0	0.03	0.02	0.00	0.00	0.01	0.04	0.05	0.01	0.03	0.03	0.01	0.01
L1	0.04	0.05	0.03	0.04	0.06	0.03	0.05	0.04	0.02	0.03	0.02	0.06
L2	0.12	0.07	0.10	0.06	0.05	0.11	0.09	0.05	0.09	0.10	0.11	0.11
L3	0.08	0.14	0.10	0.10	0.11	0.16	0.09	0.14	0.26	0.18	0.11	0.15
L4	0.19	0.32	0.16	0.08	0.12	0.21	0.12	0.13	0.10	0.13	0.20	0.21
L5	0.23	0.09	0.14	0.20	0.16	0.09	0.12	0.23	0.15	0.36	0.39	0.11
L6	0.20	0.22	0.33	0.33	0.21	0.31	0.28	0.21	0.29	0.22	0.28	0.42
L7	0.26	0.29	0.48	0.14	0.38	0.26	0.14	0.21	0.42	0.42	0.59	0.24
L8	0.21	0.24	0.30	0.22	0.33	0.50	0.28	0.36	0.61	0.43	0.38	0.32
L9	0.14	0.28	0.39	0.25	0.23	0.60	0.23	0.12	0.55	0.10	0.46	0.31
L10	0.89	0.64	0.48	0.74	0.51	0.68	0.34	0.89	0.43	0.62	0.42	0.60
L11	1.00	0.62	0.81	0.77	0.62	0.59	0.86	0.78	0.71	0.78	0.78	0.73

HIES (Ours)

	H0	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11
L0	0.01	0.06	0.35	0.31	0.13	0.09	0.07	0.07	0.04	0.07	0.40	0.22
L1	0.19	0.22	0.14	0.22	0.25	0.12	0.51	0.16	0.07	0.07	0.13	0.09
L2	0.54	0.36	0.28	0.14	0.24	0.20	0.18	0.21	0.19	0.55	0.19	0.28
L3	0.37	0.20	0.24	0.27	0.17	0.51	0.22	0.19	0.28	0.34	0.24	0.33
L4	0.32	0.22	0.33	0.36	0.24	0.44	0.36	0.45	0.35	0.29	0.37	0.33
L5	0.53	0.49	0.38	0.44	0.31	0.51	0.51	0.54	0.48	0.62	0.52	0.48
L6	0.42	0.41	0.47	0.57	0.47	0.51	0.44	0.40	0.47	0.44	0.49	0.57
L7	0.36	0.38	0.46	0.45	0.54	0.36	0.43	0.43	0.42	0.51	0.56	0.47
L8	0.30	0.30	0.40	0.26	0.34	0.47	0.40	0.26	0.39	0.34	0.36	0.33
L9	0.32	0.31	0.27	0.26	0.22	0.40	0.44	0.36	0.48	0.33	0.28	0.33
L10	0.47	0.52	0.29	0.46	0.35	0.48	0.29	0.50	0.34	0.51	0.46	0.35
L11	0.60	0.37	0.44	0.45	0.41	0.40	0.45	0.50	0.47	0.43	0.45	0.46

30%

HIS

	H0	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11
L0	0.03	0.02	0.00	0.00	0.01	0.04	0.05	0.01	0.03	0.03	0.01	0.01
L1	0.04	0.05	0.03	0.04	0.06	0.03	0.05	0.04	0.02	0.03	0.02	0.06
L2	0.12	0.07	0.10	0.06	0.05	0.11	0.09	0.05	0.09	0.10	0.11	0.11
L3	0.08	0.14	0.10	0.10	0.11	0.16	0.09	0.14	0.26	0.18	0.11	0.15
L4	0.19	0.32	0.16	0.08	0.12	0.21	0.12	0.13	0.10	0.13	0.20	0.21
L5	0.23	0.09	0.14	0.20	0.16	0.09	0.12	0.23	0.15	0.36	0.39	0.11
L6	0.20	0.22	0.33	0.33	0.21	0.31	0.28	0.21	0.29	0.22	0.28	0.42
L7	0.26	0.29	0.48	0.14	0.38	0.26	0.14	0.21	0.42	0.42	0.59	0.24
L8	0.21	0.24	0.30	0.22	0.33	0.50	0.28	0.36	0.61	0.43	0.38	0.32
L9	0.14	0.28	0.39	0.25	0.23	0.60	0.23	0.12	0.55	0.10	0.46	0.31
L10	0.89	0.64	0.48	0.74	0.51	0.68	0.34	0.89	0.43	0.62	0.42	0.60
L11	1.00	0.62	0.81	0.77	0.62	0.59	0.86	0.78	0.71	0.78	0.78	0.73

HIES (Ours)

	H0	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11
L0	0.01	0.06	0.35	0.31	0.13	0.09	0.07	0.07	0.04	0.07	0.40	0.22
L1	0.19	0.22	0.14	0.22	0.25	0.12	0.51	0.16	0.07	0.07	0.13	0.09
L2	0.54	0.36	0.28	0.14	0.24	0.20	0.18	0.21	0.19	0.55	0.19	0.28
L3	0.37	0.20	0.24	0.27	0.17	0.51	0.22	0.19	0.28	0.34	0.24	0.33
L4	0.32	0.22	0.33	0.36	0.24	0.44	0.36	0.45	0.35	0.29	0.37	0.33
L5	0.53	0.49	0.38	0.44	0.31	0.51	0.51	0.54	0.48	0.62	0.52	0.48
L6	0.42	0.41	0.47	0.57	0.47	0.51	0.44	0.40	0.47	0.44	0.49	0.57
L7	0.36	0.38	0.46	0.45	0.54	0.36	0.43	0.43	0.42	0.51	0.56	0.47
L8	0.30	0.30	0.40	0.26	0.34	0.47	0.40	0.26	0.39	0.34	0.36	0.33
L9	0.32	0.31	0.27	0.26	0.22	0.40	0.44	0.36	0.48	0.33	0.28	0.33
L10	0.47	0.52	0.29	0.46	0.35	0.48	0.29	0.50	0.34	0.51	0.46	0.35
L11	0.60	0.37	0.44	0.45	0.41	0.40	0.45	0.50	0.47	0.43	0.45	0.46

50%

HIS

	H0	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11
L0	0.03	0.02	0.00	0.00	0.01	0.04	0.05	0.01	0.03	0.03	0.01	0.01
L1	0.04	0.05	0.03	0.04	0.06	0.03	0.05	0.04	0.02	0.03	0.02	0.06
L2	0.12	0.07	0.10	0.06	0.05	0.11	0.09	0.05	0.09	0.10	0.11	0.11
L3	0.08	0.14	0.10	0.10	0.11	0.16	0.09	0.14	0.26	0.18	0.11	0.15
L4	0.19	0.32	0.16	0.08	0.12	0.21	0.12	0.13	0.10	0.13	0.20	0.21
L5	0.23	0.09	0.14	0.20	0.16	0.09	0.12	0.23	0.15	0.36	0.39	0.11
L6	0.20	0.22	0.33	0.33	0.21	0.31	0.28	0.21	0.29	0.22	0.28	0.42
L7	0.26	0.29	0.48	0.14	0.38	0.26	0.14	0.21	0.42	0.42	0.59	0.24
L8	0.21	0.24	0.30	0.22	0.33	0.50	0.28	0.36	0.61	0.43	0.38	0.32
L9	0.14	0.28	0.39	0.25	0.23	0.60	0.23	0.12	0.55	0.10	0.46	0.31
L10	0.89	0.64	0.48	0.74	0.51	0.68	0.34	0.89	0.43	0.62	0.42	0.60
L11	1.00	0.62	0.81	0.77	0.62	0.59	0.86	0.78	0.71	0.78	0.78	0.73

HIES (Ours)

	H0	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11
L0	0.01	0.06	0.35	0.31	0.13	0.09	0.07	0.07	0.04	0.07	0.40	0.22
L1	0.19	0.22	0.14	0.22	0.25	0.12	0.51	0.16	0.07	0.07	0.13	0.09
L2	0.54	0.36	0.28	0.14	0.24	0.20	0.18	0.21	0.19	0.55	0.19	0.28
L3	0.37	0.20	0.24	0.27	0.17	0.51	0.22	0.19	0.28	0.34	0.24	0.33
L4	0.32	0.22	0.33	0.36	0.24	0.44	0.36	0.45	0.35	0.29	0.37	0.33
L5	0.53	0.49	0.38	0.44	0.31	0.51	0.51	0.54	0.48	0.62	0.52	0.48
L6	0.42	0.41	0.47	0.57	0.47	0.51	0.44	0.40	0.47	0.44	0.49	0.57
L7	0.36	0.38	0.46	0.45	0.54	0.36	0.43	0.43	0.42	0.51	0.56	0.47
L8	0.30	0.30	0.40	0.26	0.34	0.47	0.40	0.26	0.39	0.34	0.36	0.33
L9	0.32	0.31	0.27	0.26	0.22	0.40	0.44	0.36	0.48	0.33	0.28	0.33
L10	0.47	0.52	0.29	0.46	0.35	0.48	0.29	0.50	0.34	0.51	0.46	0.35
L11	0.60	0.37	0.44	0.45	0.41	0.40	0.45	0.50	0.47	0.43	0.45	0.46

70%

HIS

	H0	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11
L0	0.03	0.02	0.00	0.00	0.01	0.04	0.05	0.01	0.03	0.03	0.01	0.01
L1	0.04	0.05	0.03	0.04	0.06	0.03	0.05	0.04	0.02	0.03	0.02	0.06
L2	0.12	0.07	0.10	0.06	0.05	0.11	0.09	0.05	0.09	0.10	0.11	0.11
L3	0.08	0.14	0.10	0.10	0.11	0.16	0.09	0.14	0.26	0.18	0.11	0.15
L4	0.19	0.32	0.16	0.08	0.12	0.21	0.12	0.13	0.10	0.13	0.20	0.21
L5	0.23	0.09	0.14	0.20	0.16	0.09	0.12	0.23	0.15	0.36	0.39	0.11
L6	0.20	0.22	0.33	0.33	0.21	0.31	0.28	0.21	0.29	0.22	0.28	0.42
L7	0.26	0.29	0.48	0.14	0.38	0.26	0.14	0.21	0.42	0.42	0.59	0.24
L8	0.21	0.24	0.30	0.22	0.33	0.50	0.28	0.36	0.61	0.43	0.38	0.32
L9	0.14	0.28	0.39	0.25	0.23	0.60	0.23	0.12	0.55	0.10	0.46	0.31
L10	0.89	0.64	0.48	0.74	0.51	0.68	0.34	0.89	0.43	0.62	0.42	0.60
L11	1.00	0.62	0.81	0.77	0.62	0.59	0.86	0.78	0.71	0.78	0.78	0.73

HIES (Ours)

	H0	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11
L0	0.01	0.06	0.35	0.31	0.13	0.09	0.07	0.07	0.04	0.07	0.40	0.22
L1	0.19	0.22	0.14	0.22	0.25	0.12	0.51	0.16	0.07	0.07	0.13	0.09
L2	0.54	0.36	0.28	0.14	0.24	0.20	0.18	0.21	0.19	0.55	0.19	0.28
L3	0.37	0.20	0.24	0.27	0.17	0.51	0.22	0.19	0.28	0.34	0.24	0.33
L4	0.32	0.22	0.33	0.36	0.24	0.44	0.36	0.45	0.35	0.29	0.37	0.33
L5	0.53	0.49	0.38	0.44	0.31	0.51	0.51	0.54	0.48	0.62	0.52	0.48
L6	0.42	0.41	0.47	0.57	0.47	0.51	0.44	0.40	0.47	0.44	0.49	0.57
L7	0.36	0.38	0.46	0.45	0.54	0.36	0.43	0.43	0.42	0.51	0.56	0.47
L8	0.30	0.30	0.40	0.26	0.34	0.47	0.40	0.26	0.39	0.34	0.36	0.33
L9	0.32	0.31	0.27	0.26	0.22	0.40	0.44	0.36	0.48	0.33	0.28	0.33
L10	0.47	0.52	0.29	0.46	0.35	0.48	0.29	0.50	0.34	0.51	0.46	0.35
L11	0.60	0.37	0.44	0.45	0.41	0.40	0.45	0.50	0.47	0.43	0.45	0.46

Figure 2: CoLA: heatmaps of head importance and pruning across sparsity levels. For each pruning ratio (10%, 30%, 50%, 70%), we show HIS (left) and HIES (right). Rows = layers (L0–L11); columns = heads (H0–H11). Dark/grey cells mark heads pruned at the target ratio.

C.3. Sensitivity Analysis - Ablation on α

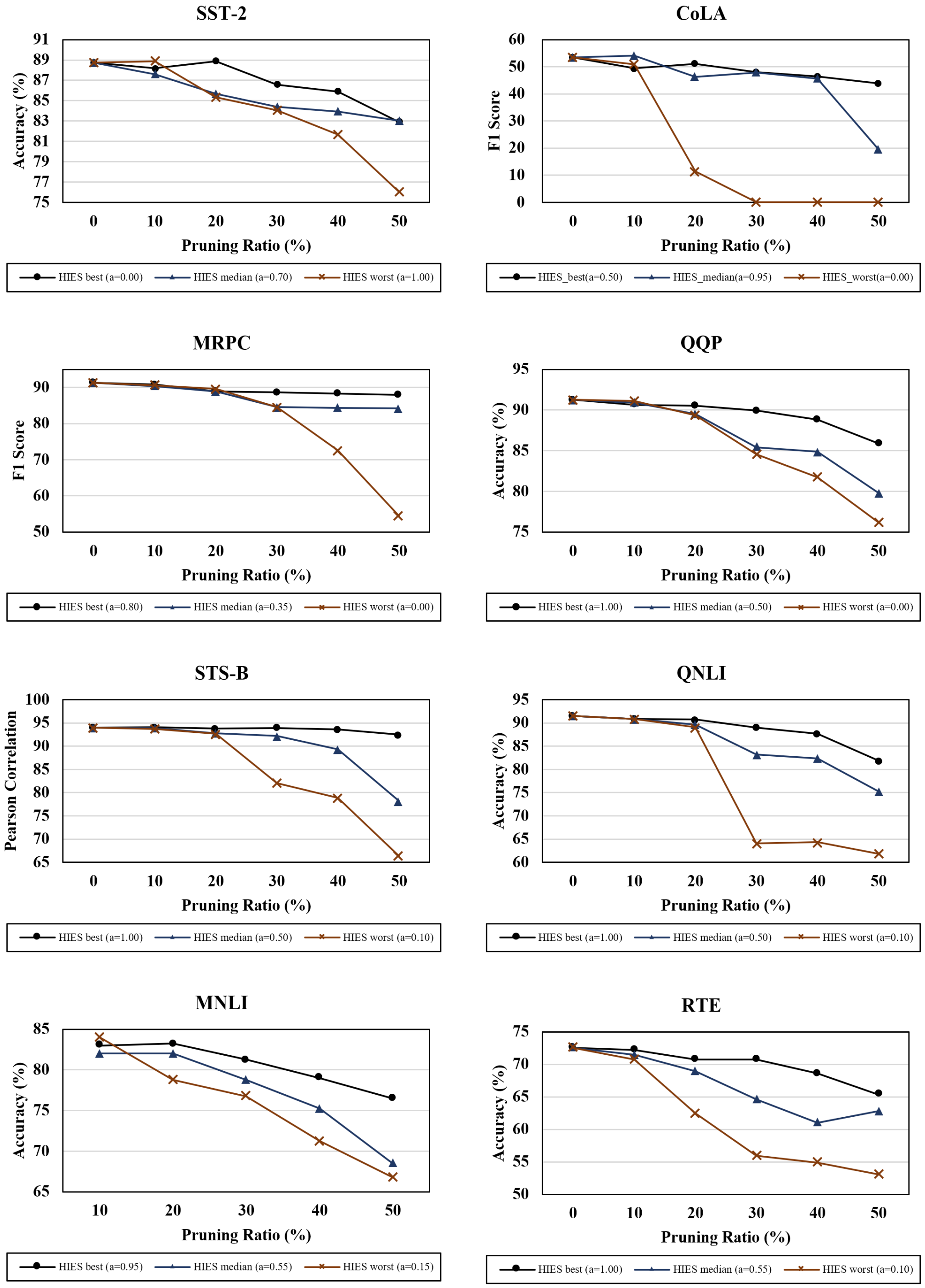


Figure 3: HIES sensitivity to the mixing coefficient α on GLUE. For each task, we sweep α and report three choices—α_{best}, α_{median}, α_{worst}—selected by weighted AUC (wAUC) across pruning ratios. Curves plot performance versus pruning ratio for these three settings.

We sweep the mixing coefficient $\alpha \in [0,1)$ that interpolates the gradient-based head-importance (HIS) and attention-entropy (AE) signals in HIES,

$$\mathrm{HIES}_h(\alpha) = \alpha\,\widehat{\mathrm{HIS}}_h + (1-\alpha)\,\widehat{\mathrm{AE}}_h.$$

As expected, larger α upweights HIS and preserves heads with strong task relevance, whereas smaller α upweights AE and retains low-entropy, focused heads. We choose a single $\alpha^\star$ on a held-out validation split and fix it for all reported experiments; the resulting accuracy–sparsity profiles are shown in Figure 3.

C.4. Results on LLaMA2-7B

On LLaMA-2-7B, we evaluate zero-shot pruning on **HellaSwag**, **ARC-e**, and **ARC-c**, comparing HIES with HIS across pruning ratios 10–60%. On average, HIES attains higher accuracy and stability than HIS and this advantage persists uniformly, indicating that the same pruning mechanism scales to larger models with higher head counts.

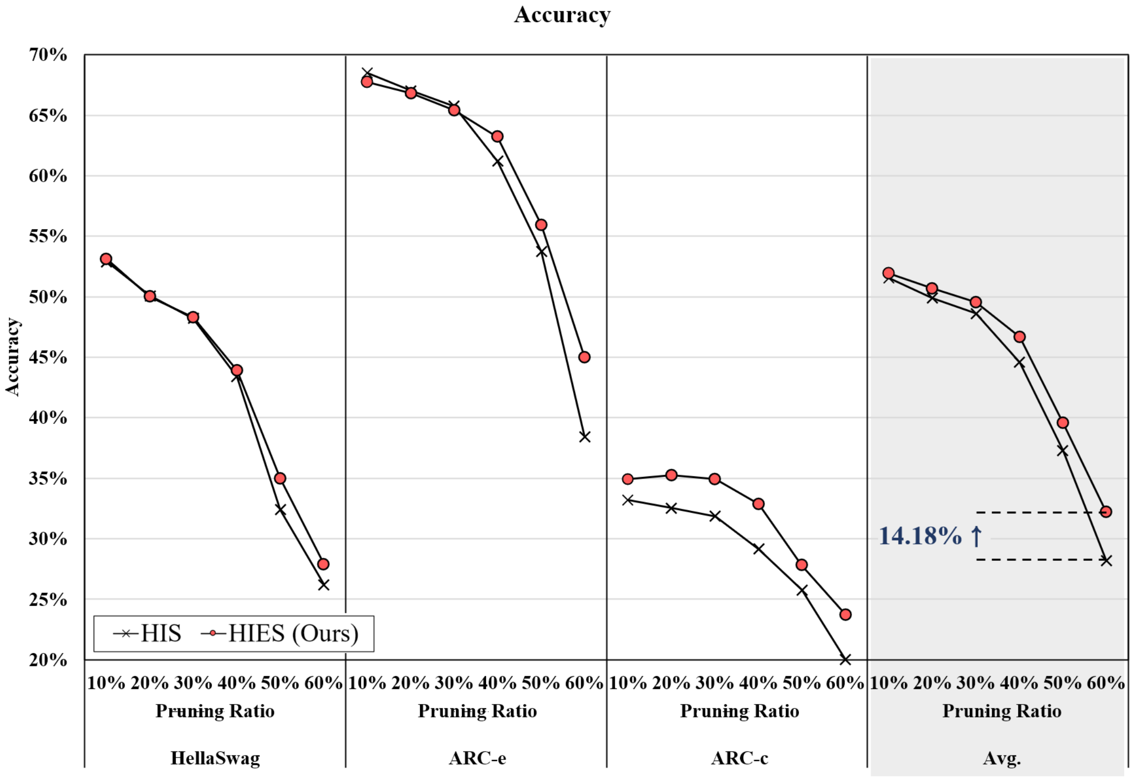


Figure 4: **Zero-shot performance on LLaMA-2-7B — Accuracy.** Comparison of HIES (ours) and HIS on **HellaSwag**, **ARC-e**, and **ARC-c** across pruning ratios 10–60% (x-axis). Each panel reports task *accuracy*, and the rightmost panel reports the mean over tasks. HIES matches or exceeds HIS throughout, yielding higher average accuracy across sparsity levels.

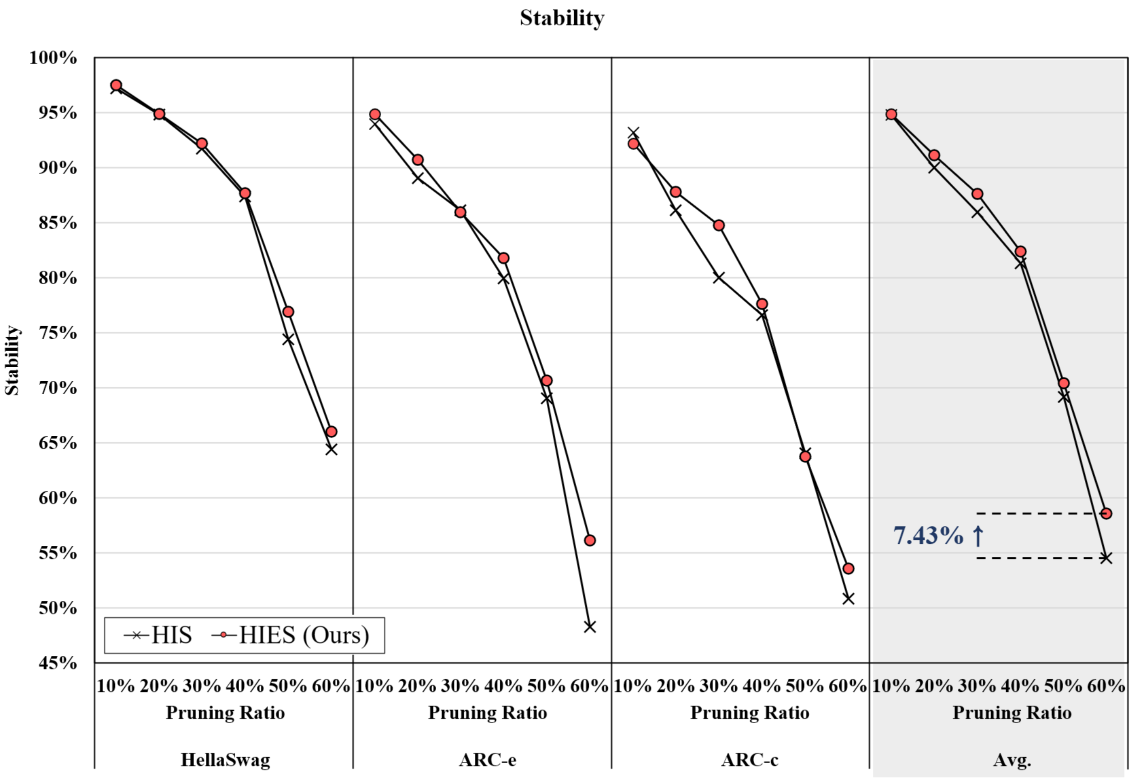


Figure 5: **Zero-shot performance on LLaMA-2-7B — Stability.** Comparison of HIES (ours) and HIS on **HellaSwag**, **ARC-e**, and **ARC-c** across pruning ratios 10–60% (x-axis). Each panel reports task *stability*, and the rightmost panel shows the task average. HIES consistently improves stability over HIS across pruning ratios, indicating scalability to larger models.