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

Anonymous ACL submission

Abstract

Transformer-based models have achieved remarkable performance. However, their structural characteristics—multiple layers and attention heads—introduce efficiency 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 15.2% improvement in model quality and 2.04× improvement in stability over HIS-only methods, enabling substantial model compression without sacrificing either accuracy or stability. Code will be released upon publication.

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 (Kaplan et al., 2020; Chen et al., 2025b). This scaling, however, incurs substantial computational and memory costs during inference, resulting in prohibitive latency and energy consumption (Yang et al., 2020; Kim and Wu, 2020; Hoefler et al., 2021; Zhou et al., 2024). These constraints become critical barriers 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 assistants, and personalized recommendation systems.

To improve deployability of LLMs for resource-constrained environments, various pruning methods have been proposed (Ma et al., 2023; Yang et al., 2024). Typically, 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 important heads based on Head Importance Score (HIS), which quantifies the gradient-based contribution of each head to the loss function. By leveraging gradient-based sensitivity to the loss, HIS prioritizes heads that have the most direct impact on accuracy of model inference.

However, HIS-based methods often exhibit limited stability in their performance. For clarity, prior works (Bair et al., 2023; Blanchet et al., 2024) motivate treating stability as a practical surrogate for robustness—namely, a model's resilience to input perturbations and pruning-induced distributional shifts. Such stability is crucial in real-world deployments where distribution shifts are common and aggressive compression is often required. In our observations, this instability appears to stem from two key factors. First, existing HIS-based methods solely rely on the loss gradient with respect to each head's output, which fails to capture token-level attention allocation or its alignment with the task's empirical distribution. Consequently, a concentrated head and a diffuse head can yield similar important scores, concealing their functionally distinct roles on the task-specific data manifold. Second, a uniform, layer-agnostic criterion precludes layer-specific adaptation despite evidence that different layers require distinct attention behaviors (Artzy and Schwartz, 2024). Lacking such layer-specific characteristics often results in imbalanced prun-

ing—preserving redundant heads in some layers while removing functionally critical ones in others. This imbalance not only degrades accuracy but also undermines stability, leading to unpredictable performance fluctuations across inputs or compression levels, particularly under aggressive pruning ratios.

This work aims to address the aforementioned limitations by proposing an *Entropy-Aware Pruning Criterion*, termed **HIES (Head Importance-Entropy Score)**, which jointly 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. 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 principled combination in HIES enables layer-adaptive pruning decisions and preserves functionally important heads. This allows for more balanced pruning across layers, improving both accuracy and stability under aggressive compression. Empirically, HIES yields up to a 15.2% improvement in model quality and 2.04× improvement in stability over HIS-only methods. By preserving 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

Attention head pruning. To compress large language models efficiently, structured pruning methods (Han et al., 2015; Wang et al., 2019; Xia et al., 2022; Ma et al., 2023; Ashkboos et al., 2024), 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 reduces attention FLOPs and KV-cache memory 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 (Jaradat et al., 2024; Muralidharan et al., 2024)[1]. Attention head pruning removes selected heads from a trained Transformer's multi-head attention with minimal impact on end-task performance (Vaswani et al., 2017). A widely adopted criterion is the HIS of Michel et al. (2019), which introduces mask variables $m_h \in \{0, 1\}$ multiplying the output of head h and defines importance as the expected first-order loss increase under masking:

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

where $\mathcal{D}$ denotes an input sample drawn from the data distribution $\mathcal{D}$, $\mathcal{L}(x)$ 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 HIS_h and pruned in ascending order of importance.

Attention Entropy and Stability. Zhai et al. (2023) quantify the concentration of each attention head's focus over input tokens via the entropy of its attention weight distribution $\mathrm{AE}_h = (H(p^{(h)}) = -\sum_{i=1}^{n} p_i^{(h)} \log p_i^{(h)}$, 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, including oscillations in the loss landscape and even divergence across various model scales and tasks. During inference, pruning introduces structured perturbations to the original attention mechanism, which can lead to instability in model predictions. To quantify this inference stability, we measure the prediction consistency between the pruned and original models, following the stability framework of Bousquet and Elisseeff (2002): Stability = $\frac{1}{|\mathcal{D}|}\sum_{x\in\mathcal{D}} \mathbf{1}\left[y_{\text{pruned}}(x) = y_{\text{original}}(x)\right]$.

3 Motivation

Pruning Transformer models is most commonly driven by gradient-based criteria, such as HIS and variants used in recent pruning frameworks (e.g., LLM-Pruner) (Michel et al., 2019; Ma et al., 2023). While gradient-based methods are often effective at moderate sparsity, they exhibit sharp accuracy degradation once the pruning ratio exceeds a certain threshold, as shown in Figure 1 (a). Such *sharp-drops* have been widely observed across various attention variants and tasks (Ma et al., 2023; Li et al., 2023a; Ghattas et al., 2025), underscoring the generality of this phenomenon.

[1]For more detailed discussions on related work, please refer to Appendix A.

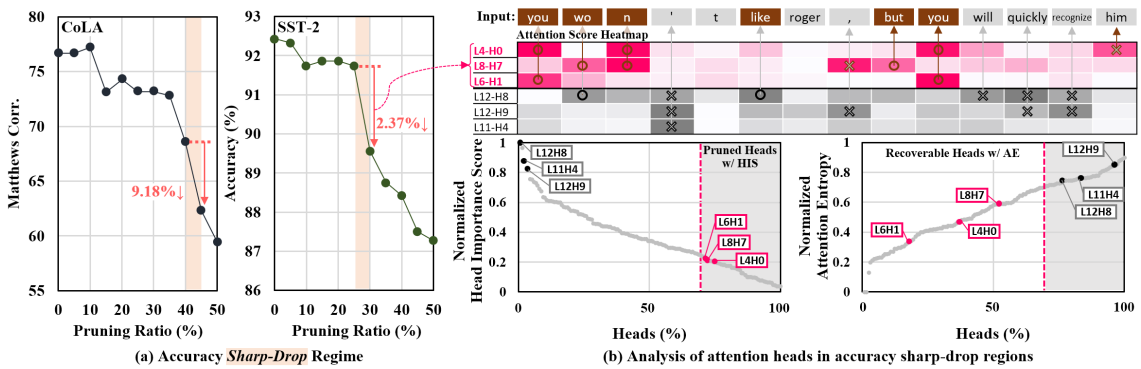


Figure 1: Analysis of accuracy degradation and head behaviors under HIS-based pruning. In our diagnostic study, we analyze the phenomena of pruning by HIS on BERT, focusing on detailed attention head behaviors during inference. (a) Accuracy curves of HIS-based pruning on CoLA and SST-2. The bold color highlights the sharp-drop regime of HIS-based pruning. (b) Head-level analysis on SST-2 with a validation example misclassified by the HIS-pruned model. The attention score heatmap shows heads on each column, where pruned low-HIS heads are indicated with colored layer–head labels and unpruned high-HIS heads with gray table. It then shows token-wise distributions, where tokens deemed important for classification (e.g., sentiment-discriminative tokens in SST-2) are marked with **O**, and non-critical tokens with **X**. The left plot shows the distribution of heads by normalized HIS, and the right plot shows the distribution by normalized AE, where pruned low-HIS heads are highlighted with red boxes and unpruned high-HIS heads with gray boxes. See Appendix C for experimental setup details.

We focus on this "*sharp-drop*" regime and contrast two groups of heads. The first group consists of low-HIS heads that are pruned during the sharp-drop of accuracy. The second group consists of high-HIS heads that remain unpruned. The attention score heatmap in Figure 1 (b) reveals that some pruned heads (red table) assign high attention scores—i.e., the weights computed by the softmax over token-token similarity that indicate how strongly a token attends to another—to sentiment-discriminative tokens (tokens relevant for classification). Nonetheless, these heads are pruned due to their low HIS values and end up causing the sharp accuracy drop observed in Figure 1 (a). In contrast, some unpruned heads (gray table) often allocate strong attention to non-informative tokens. These heads, however, have high HIS values and thus remain unpruned, though they contribute little to overall model quality. These analysis results demonstrate that the gradient-based HIS is insufficient to capture the token-level attention score distributions (the detailed mathematical analysis is provided in Section 4.2.1), thereby resulting in suboptimal pruning decisions for heads focusing on decisive tokens.

The *sharp-drop* observed in pruning can be interpreted as a collapse of structural diversity in attention behaviors, caused by the elimination of heads that concentrate on decisive tokens. To capture and prevent such collapse of structural diversity in attention, we employ attention entropy, a measure widely used to prevent policy collapse in reinforcement learning (Bharadhwaj et al., 2020; Xiao et al., 2021; Wang et al., 2025). As shown in the bottom-right plots of Figure 1 (b), incorporating AE helps retain low-entropy heads by recognizing their concentrated focus on decisive tokens, thereby preventing them from being pruned and mitigating the *sharp-drop* in accuracy.

Building on our analysis, we posit that attention entropy captures structural signals that reflect unstable behavior during deployment, leading to the following hypothesis:

Attention entropy serves as an indicator of inference-time stability, mitigating accuracy sharp-drops

In particular, low-entropy heads may correlate with increased sensitivity to input perturbations, leading to unstable predictions under distribution shifts. This perspective motivates our investigation of entropy as a proxy for robustness and consistency during inference.

4 Proposed Method

Figure 2 provides an overview of our proposed *Head Importance–Entropy Score (HIES)*. Figure 2 (a) outlines the Transformer architecture, where HIS computes the importance score of each of the h attention heads. Figure 2 (b) displays the heatmap of the first-order, loss-based HIS. Combining the normalized HIS and AE yields the HIES heatmap in Figure 2 (c), which integrates complementary signals provide a more stable assessment of atten-

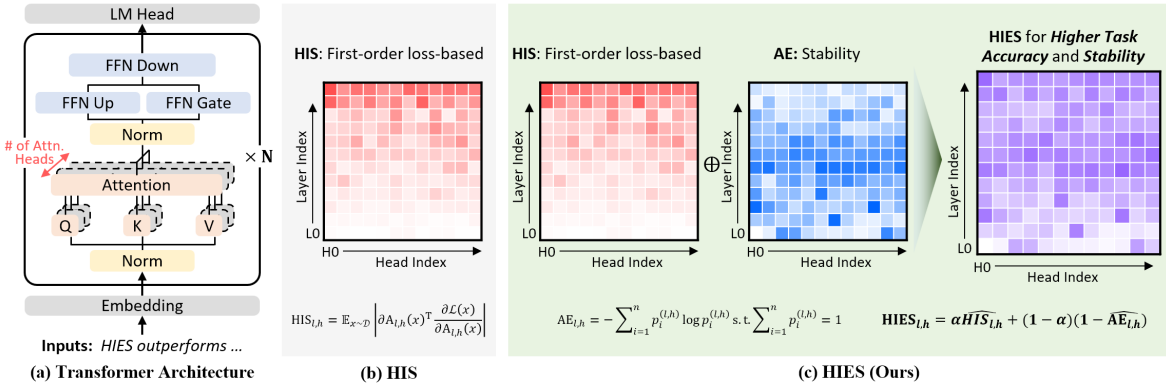


Figure 2: Design overview of Head Importance-Entropy Score (HIES). Darker cells correspond to values closer to 1, while lighter cells correspond to values closer to 0. Note the heatmap utilizes each metric across attention heads in BERT on the CoLA dataset.

tion heads across layers. This design captures the key intuition of HIES: HIS can be reinforced by AE, providing a more robust basis for pruning head selections. Section 4.1 formalizes HIES, and Section 4.2 develops a risk-decomposition analysis that clarifies the complementary roles of HIS and AE and motivates HIES as a robust criterion for head selection in pruning.

4.1 Head Importance Entropy Score

We define the **Head Importance–Entropy Score (HIES)** as a *linear combination* of the normalized HIS and AE deficit:

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

where $\alpha \in [0, 1)$ is a tunable mixing coefficient[2].

This form arises directly from the risk decomposition developed in Section 4.2. Pruning trades off two complementary objectives: minimizing loss sensitivity, captured by a first-order gradient score (Molchanov et al., 2019), and preserving specialized, low-entropy heads, captured by $1 - \mathrm{AE}$. A weighted-sum scalarization of the two recovers the full set of Pareto-optimal trade-offs under convex conditions (Boyd and Vandenberghe, 2004), and the two terms map cleanly onto our analysis: the loss-increase bound (Lemma 4.1) motivates $\widehat{\mathrm{HIS}}_h$, while the stability-based generalization-gap bound (Eq. (5); Bousquet and Elisseeff, 2002) motivates $1 - \widehat{\mathrm{AE}}_h$. Their gradient directions are orthogonal in expectation (Lemma 4.3), so the two signals combine linearly without gradient conflict (Yu et al., 2020), and Lemma 4.2 shows that greedy selection by HIES is globally optimal for the cardinality-constrained head-selection problem.

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 signals. 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 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 (De Amorim et al., 2023; Lima and Souza, 2023).

4.2 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 token-wise 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 retaining heads with high HIES, we simultaneously minimize our theoretical bound and enhance pruning stability. Conceptually, this analysis formalizes importance-based selection into a principled framework and offers a rigorous rationale for HIES's safety and effectiveness.

4.2.1 Loss-Increase Control via Head Importance (HIS)

Setup. Let n be the sequence length, $|\mathrm{H}|$ the number of heads, d_v the value dimension per head,

[2]For the main results we adopt task-specific tuning of α; sensitivity is analyzed in Appendix D.8, and a robustness check under a single fixed α across model families is reported in Appendix D.9.

and $d = |\mathrm{H}|\, d_v$ the model width (i.e., hidden size). For head h, let $A_h \in \mathbb{R}^{n \times d_v}$ denote the head output, i.e., the value-projected attention representation $A_h = \mathrm{softmax}\left(\frac{Q_h K_h^\top}{\sqrt{d_k}}\right) V_h$, where $Q_h = XW_h^Q \in \mathbb{R}^{n \times d_k}$, $K_h = XW_h^K \in \mathbb{R}^{n \times d_k}$, and $V_h = XW_h^V \in \mathbb{R}^{n \times d_v}$ are the query, key, and value projections of the input $X \in \mathbb{R}^{n \times d_{\text{model}}}$, with parameter matrices $W_h^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_h^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, and $W_h^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$.

We then define $\mathbf{y} = \mathrm{Concat}(A_1, \ldots, A_{|\mathrm{H}|}) \in \mathbb{R}^{n \times d}$ as the pre-projection representation, which is subsequently projected through $W^O \in \mathbb{R}^{d \times d}$. Head removal is modeled by mask variables $m_h \in \{0, 1\}$: $\delta A_h = -(1 - m_h) A_h$ and $\delta \mathbf{y} = \mathrm{Concat}(\delta A_1, \ldots, \delta A_{|\mathrm{H}|})$. Formal definitions and implementation notes are deferred to Appendix B.1.2.

Head Importance Score (HIS). We define

$$\begin{aligned} \mathrm{HIS}_h &:= \mathbb{E}_{x \sim \mathcal{D}} \left| \frac{\partial \mathcal{L}(x)}{\partial m_h} \right| = \mathbb{E}_{x \sim \mathcal{D}} \left| \left\langle \nabla_{A_h(x)} \mathcal{L}(x),\, A_h(x) \right\rangle_F \right| \\ &= \mathbb{E}_{\mathcal{D}} \Big[\left| \langle \nabla_{A_h} \mathcal{L},\, A_h \rangle_F \right| \Big]. \end{aligned} \tag{2}$$

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 B.1.6).

Lemma 4.1 (Loss-increase upper bound under head masking)**.** *Let $\beta_y := \|\nabla_{\mathbf{y}}^2 \mathcal{L}\|_2$. For any mask variables $\{m_h\}_{h=1}^{|\mathrm{H}|}$,*

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

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

$$\Delta\mathcal{L} \le \sum_{h=1}^{|H|} (1 - m_h)\, \mathrm{HIS}_h + \frac{1}{8} \|W^O\|_2^2 \sum_{h=1}^{|H|} (1 - m_h)\, \|A_h\|_F^2. \tag{3}$$

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

Implication for pruning. Eq. (3) shows that, for a fixed pruning fraction $\rho = \frac{1}{|\mathrm{H}|} \sum_h (1 - m_h)$, 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 B.1.5), justifying the use of HIS_h as a practical surrogate importance for head pruning.

However, since $A_h = \mathrm{Attn}_h V_h$, HIS_h in Eq. (2) depends solely on A_h and $\nabla_{A_h} \mathcal{L}$, without explicitly incorporating the distribution Attn_h. Consequently, two heads with very different attention patterns (e.g., sharply focusing on tokens versus spreading over tokens) can yield similar HIS_h whenever the resulting A_h is comparable. This is supported by our empirical analysis in Section 3.

4.2.2 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 (Bousquet and Elisseeff, 2002; Shalev-Shwartz and Ben-David, 2014; Hardt et al., 2016).

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

$$\begin{aligned} \mathrm{AD}_h(x) &:= \frac{1}{n(x)\, \log n(x)} \times \sum_{t=1}^{n(x)} \Big(\log n(x) - H\big(\boldsymbol{\alpha}_t^{(h)}(x)\big) \Big) \\ &= 1 - \mathrm{AE}_h(x) \in [0, 1]. \end{aligned} \tag{4}$$

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

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

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 B.2.

4.2.3 Risk Upper Bound and HIES Minimization

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

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 Acc.	CoLA Matthews corr	MRPC F1 Score	QQP Acc.	STS-B Pearson corr	QNLI Acc.	MNLI Acc.	RTE Acc.	Average	
0%	$\text{BERT}_{\text{base}}$	92.43	76.69	91.35	91.27	94.02	91.54	84.57	72.56	86.80	
10%	Random	92.09	75.25	89.90	91.00	93.88	89.87	82.62	70.76	85.67	0.00%
	AD	92.55	75.48	90.56	91.10	93.65	90.87	83.50	70.40	86.01	+0.40%
	HIS	91.74	**77.21**	90.65	91.23	**94.04**	90.63	**84.04**	**71.84**	86.42	+0.88%
	L2	90.37	74.00	83.78	69.23	83.12	60.81	67.75	49.82	72.36	-15.54%
	LLM-Pruner (Channel)	91.97	76.05	79.19	83.53	93.06	87.39	80.46	67.51	82.40	-3.82%
	LLM-Pruner (Block)	91.06	76.69	89.83	84.96	93.77	87.42	82.03	64.62	83.80	-2.19%
	SliceGPT (w/o tune)	51.38	49.86	81.46	63.18	58.92	53.91	36.84	49.46	55.63	-35.07%
	SliceGPT (w/ tune)	86.47	61.87	82.30	88.28	62.47	83.45	77.34	54.51	74.59	-12.94%
	HIES (ours)	**92.66**	75.48	**91.04**	**91.93**	94.00	**91.03**	**84.04**	**71.84**	**86.50**	**+0.97%**
30%	Random	90.29	69.02	85.27	84.00	92.90	79.80	75.15	60.29	79.59	0.00%
	AD	86.58	50.00	84.53	84.57	81.94	67.82	77.00	56.68	73.64	-7.47%
	HIS	89.56	73.17	**89.37**	89.95	93.82	89.04	82.25	68.59	84.47	+6.13%
	L2	86.58	67.52	81.58	64.83	76.52	51.09	56.00	50.90	66.88	-15.97%
	LLM-Pruner (Channel)	88.53	70.36	86.89	81.23	92.50	67.38	66.67	64.26	77.23	-2.97%
	LLM-Pruner (Block)	88.99	73.93	84.76	80.09	93.40	82.68	78.33	66.79	81.12	+1.92%
	SliceGPT (w/o tune)	50.80	53.29	78.05	63.18	54.64	53.18	34.90	51.99	55.00	-30.90%
	SliceGPT (w/ tune)	83.49	60.14	81.80	85.80	60.89	67.36	75.50	54.87	71.23	-10.49%
	HIES (ours)	**91.86**	**74.97**	88.81	**90.37**	**93.89**	**89.13**	**82.50**	**70.04**	**85.20**	**+7.04%**
50%	Random	78.74	61.02	72.53	66.25	91.40	67.53	67.32	53.79	69.82	0.00%
	AD	82.91	50.00	54.50	76.18	75.00	68.94	68.00	55.96	66.44	-4.84%
	HIS	87.27	59.48	86.52	**85.91**	92.61	**82.68**	78.67	62.82	79.50	+13.84%
	L2	82.80	60.98	85.30	64.83	69.76	50.54	44.42	47.29	63.24	-9.39%
	LLM-Pruner (Channel)	86.47	61.64	83.92	81.47	89.74	60.66	67.42	62.82	74.27	+6.37%
	LLM-Pruner (Block)	87.84	**70.09**	83.84	78.80	**92.69**	72.60	73.60	61.73	77.65	+11.20%
	SliceGPT (w/o tune)	50.92	52.79	81.22	63.19	47.70	50.98	34.93	50.90	54.08	-22.54%
	SliceGPT (w/ tune)	83.49	57.45	81.37	82.16	55.05	65.79	71.70	51.62	68.58	-1.78%
	HIES (ours)	**90.71**	68.52	**86.80**	85.73	92.65	**82.68**	**79.00**	**65.34**	**81.43**	**+16.63%**

$$\mathcal{R}(m) \;:=\; \sum_{h=1}^{|\mathrm{H}|} (1 - m_h)\,\mathrm{HIES}_h \tag{6}$$

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

$$\min_{m \in \{0,1\}^{|\mathrm{H}|}} \sum_{h=1}^{|\mathrm{H}|} (1 - m_h)\,\mathrm{HIES}_h \quad \text{s.t.} \sum_{h=1}^{|\mathrm{H}|} m_h \;=\; k. \tag{7}$$

Lemma 4.2 (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 (6) subject to (7).*

4.2.4 Orthogonality and Complementarity

Lemma 4.3 (Orthogonality)**.** *Let $u_h := \mathrm{sign}(\boldsymbol{\alpha}^{(h)\top} g_h)\, g_h, v_h := \mathbf{1} + \log \boldsymbol{\alpha}^{(h)}, \tilde{u}_h := P\,u_h, \quad \tilde{v}_h := P\,v_h$, 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 $\langle \mathbb{E}_{\mathcal{S}}[\tilde{u}_h], \mathbb{E}_{\mathcal{S}}[\tilde{v}_h] \rangle = 0$[3], then*

$$\mathbb{E}_{\mathcal{S}}\Big[\big\langle \widetilde{\nabla}_{\boldsymbol{\alpha}^{(h)}} \mathrm{HIS}_h,\, \widetilde{\nabla}_{\boldsymbol{\alpha}^{(h)}} \mathrm{AE}_h \big\rangle\Big] \;=\; 0, \tag{8}$$

i.e., the two gradient directions are orthogonal in expectation[4].

[3]We provide an empirical sanity check supporting the assumption, which is consistently validated on both $\text{BERT}_{\text{base}}$ and $\text{LLaMA-2}_{\text{7B}}$ in Appendix D.1.

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

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 distributional concentration. Combined, they balance pruning minimally influential heads and preserving heads important for generalization, underpinning HIES's effectiveness.

5 Experimental Results

5.1 Experimental Setup

Model and Dataset. We extensively evaluate HIES across 9 model architectures and 21 datasets spanning language, vision, and multimodal tasks; detailed model and dataset descriptions are provided in Appendix C.2.

Baselines. We compare HIES against seven representative pruning methods spanning three granularities. At the head level we consider **Random** (uniform head removal), **L2-Norm** (heads with smaller ℓ_2 weight magnitudes are pruned first), **HIS** (Michel et al., 2019) (first-order gradient-based head importance, pruned in ascending order), and **Attention Deficit** (**AD**; 1−Attention Entropy) (Zhai et al., 2023) (heads with smaller attention entropy, i.e., more concentrated patterns, are pruned first). At the channel and block level, **LLM-Pruner** (Ma et al., 2023) provides two variants restricted to attention

Table 2: Stability results with $\text{BERT}_{\text{base}}$ on GLUE tasks. We report stability (%) against the unpruned model. Percentage improvements (in blue) are relative to HIS within each pruning ratio.

Pruning Ratio	Method	SST-2	CoLA	MRPC	QQP	STS-B	QNLI	MNLI	RTE	Average	
10%	HIS	97.71	**96.55**	**94.36**	97.30	**99.47**	95.79	95.66	94.22	96.38	0.00%
	L2	95.18	76.41	26.72	68.63	21.87	60.33	72.44	39.71	57.66	-40.17%
	AD	**98.51**	96.36	93.63	**97.80**	89.07	**97.25**	**96.15**	**94.95**	95.47	-0.94%
	LLM-Pruner	97.02	85.81	75.25	85.40	39.53	90.39	84.98	86.28	80.58	-16.39%
	HIES (ours)	98.17	96.36	94.12	**97.80**	96.67	**97.25**	95.74	93.50	**96.20**	**-0.19%**
30%	HIS	94.38	90.03	**90.03**	94.77	78.93	92.95	**91.56**	87.36	90.00	0.00%
	L2	90.94	83.13	26.72	63.90	20.73	50.39	59.23	42.24	54.66	-39.27%
	AD	88.53	81.21	78.68	87.90	24.53	63.04	79.07	69.68	71.58	-20.47%
	LLM-Pruner	92.66	63.09	84.80	82.37	62.13	68.70	68.29	70.04	74.01	-17.77%
	HIES (ours)	**97.13**	**93.38**	89.46	**94.97**	**86.67**	**93.23**	**91.56**	**88.81**	**91.90**	**+2.11%**
50%	HIS	90.02	40.84	83.33	**88.80**	69.13	**85.54**	**81.47**	78.70	77.23	0.00%
	L2	86.47	82.26	26.72	63.90	20.40	49.81	45.58	34.30	51.18	-33.73%
	AD	52.18	81.21	51.72	72.33	21.87	73.49	68.49	62.82	60.51	-21.65%
	LLM-Pruner	89.68	41.32	78.19	83.27	44.13	61.58	67.52	75.81	67.69	-12.35%
	HIES (ours)	**95.07**	**83.13**	**84.56**	88.27	**75.20**	**85.54**	78.29	**81.23**	**83.91**	**+8.65%**

layers: a channel-wise version that prunes attention channels by first-order Taylor importance, and a block-wise version that applies the same criterion at attention-block granularity while retaining the first three and final layers, as in prior work. At the subspace level, **SliceGPT** (Ashkboos et al., 2024) projects activations onto principal components estimated from calibration data and removes directions corresponding to less-important subspaces.

5.2 Main Results

We evaluate HIES using two key metrics: model quality and stability[5]. As reported in Table 1, HIES improves model quality by 8.21% on average. Table 2 further demonstrates a 3.3% average stability gain over HIS. Notably, at a pruning ratio of 50%, HIES achieves gains of up to 15.2% in model quality and 2.04× in stability compared to the best-performing baseline. These results corroborate our theoretical analysis, demonstrating that HIES preserves end-task performance while markedly enhancing robustness, both critical for reliable and efficient deployment.

5.3 Extended Experimental Results

5.3.1 Head Removal Patterns (Heatmap)

In our main results, HIES exhibits more stable performance gains than HIS at aggressive pruning ratios (e.g., ≥,30%). At lower ratios (e.g., ≤,10%), HIS and HIES perform comparably, while HIS achieves marginally higher accuracy in a few cases. We posit distinct prioritization. HIS, a gradient-based metric tends to retain redundant, low-risk heads at low sparsity, while HIES adds attention entropy to capture stability, thereby preserving specialized low-entropy heads that enhance robustness.

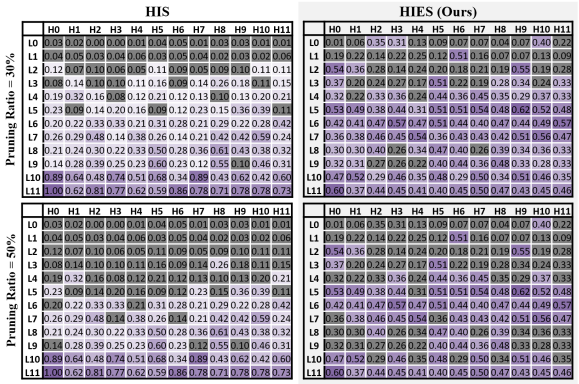


Figure 3: Head-level pruning patterns on the CoLA dataset across pruning ratios from 30% to 50% pruning ratios. Pruned heads are shaded in gray.

Pruning heatmap analysis provides empirical support for this distinction. As shown in Figure 3, HIS (left panels) tends to remove heads primarily from the lower layers, producing an approximately bottom-up pattern. This behavior is guided by a one-step gradient saliency, which estimates the importance of each attention head based on a single backward pass through the model— this assigns higher importance to heads whose activations have a larger immediate effect on the loss. Meanwhile, HIES (right panels) yields a more dispersed selection spanning lower, middle, and upper layers — other tasks also exhibit similar patterns (details are provided in Appendix D.4). 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. Consequently, HIES exhibits more stable performance across pruning ratios, yielding flatter accuracy–sparsity curves than HIS.

[5]Experimental details are provided in Appendix C. We also conduct sensitivity analysis across multiple random seeds and observe consistent improvements for both $\text{BERT}_{\text{base}}$ and $\text{LLaMA-2}_{\text{7B}}$. Details are provided in Appendix D.2. Note that we use a calibration size of 32 for computing HIES; additional analyses are reported in Appendix D.3.

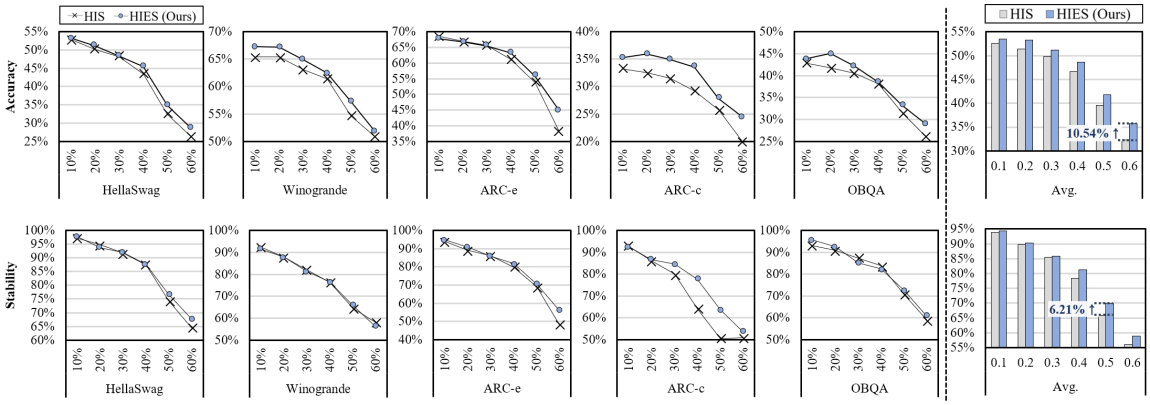


Figure 4: Comparison of HIES (Ours) and HIS on HellaSwag, Winogrande, ARC-e, ARC-c, and OBQA across pruning ratios from 10% to 60% (x-axis). The top row reports task accuracy, while the bottom row reports stability. The rightmost panels summarize the mean over tasks.

Table 3: Experimental results with $\text{ViT}_{\text{Large}}$ on image classification benchmarks and $\text{LLaVA-1.5}_{\text{Large}}$ on multi-modal tasks. Relative improvements of HIES (Ours) over HIS are shown in blue.

	Vision Transformer Model ($\text{ViT}_{\text{Large}}$)							Vision-Langauge Model ($\text{LLaVA-1.5}_{\text{7B}}$)						
	Image Classification								Visual Question Answering		Complex Multimodal			
	ImageNet1k		CIFAR-100		Avg.				VizWiz-VQA		MM-Vet		Avg.	
	HIS	**HIES (Ours)**	HIS	**HIES (Ours)**	HIS	**HIES (Ours)**			HIS	**HIES (Ours)**	HIS	**HIES (Ours)**	HIS	**HIES (Ours)**
10%	**86.40%**	84.40%	91.40%	**92.40%**	**88.90%**	88.40% -0.56%		10%	**48.00%**	47.33%	**41.20%**	39.40%	**44.60%**	43.37% -2.84%
20%	44.80%	**80.20%**	65.60%	**85.20%**	55.20%	**82.70%** 49.82%		30%	44.00%	**47.00%**	29.80%	**34.00%**	36.90%	**40.50%** 8.89%
30%	6.20%	**55.40%**	19.40%	**40.00%**	12.80%	**47.70%** 256.66%		50%	32.67%	**39.67%**	24.60%	**25.20%**	28.64%	**32.43%** 11.71%

5.3.2 Scalability to Larger Models

On $\text{LLaMA-2}_{\text{7B}}$, we evaluate pruning on HellaSwag, Winogrande, ARC-e, ARC-c, and OBQA, comparing HIES with HIS across pruning ratios of 10–60%. HIES improves accuracy by up to +10.54% and stability by up to +6.21% relative to HIS, averaged over tasks. This advantage persists uniformly across pruning ratios, indicating that the same pruning mechanism scales effectively to larger models with higher head counts. Notably, ARC-c is the most difficult benchmark in this suite, as reflected by its lowest base accuracy. Even under this challenging setting, HIES achieves consistent and often larger accuracy gains relative to HIS, underscoring its robustness not only on easier tasks but also on the hardest ones.

5.3.3 Generalizability Across Architectures and Tasks

Table 3 evaluates the generalizability of HIES across attention variants and task domains, spanning vision classification with $\text{ViT}_{\text{Large}}$ and visual-language reasoning with $\text{LLaVA-1.5}_{\text{7B}}$.

First, across attention-based variants, HIES consistently shifts the sharp accuracy-drop region to more aggressive sparsity levels: on $\text{ViT}_{\text{Large}}$ at 20% sparsity, HIES achieves 49.82% higher accuracy than HIS (vs. 55.20%).

Second, the gains carry over to multi-modal tasks: on VizWiz-VQA and MM-Vet, HIES improves accuracy by 11.71% at 50% sparsity over HIS.

Third, HIES continues to outperform HIS on large-scale reasoning (Appendix D.6) and downstream image tasks (Appendix D.7).

Finally, to verify generality across recent decoder-only architectures, we additionally evaluate HIES on seven models from the Qwen2.5 (Qwen et al., 2025), Qwen3 (Yang et al., 2025), LLaMA-3.1 (Grattafiori et al., 2024), DeepSeek-R1 (Guo et al., 2025), and OLMo-3 (Olmo et al., 2025) families, spanning Grouped-Query Attention (Ainslie et al., 2023) and standard MHA with parameter sizes from 1.7B to 8B. Using a *single fixed* α per model (no per-task tuning), HIES outperforms HIS at every pruning ratio across all seven models. Detailed results are reported in Appendix D.9.

6 Conclusion

In this paper, we presented HIES, a pruning criterion that combines gradient-based head importance with attention entropy to capture both loss sensitivity and structural specialization of attention heads. Across diverse Transformer architectures and tasks, HIES consistently outperforms baselines in accuracy and inference-time stability. We believe HIES offers a principled direction for efficient pruning of Transformer-based models, extendable to broader structured sparsity and large-scale deployment.

Limitations

While HIES yields consistent improvements across diverse Transformer architectures and tasks, two scope extensions remain for future work. First, our evaluation focuses on standard multi-head and grouped-query attention. HIES depends only on per-query-head attention probabilities and gradients, which remain well-defined under expert routing. We therefore expect it to transfer to mixture-of-experts (MoE) Transformers with minimal modification, though empirical verification is left for future work. Second, HIES is in principle orthogonal to other compression techniques such as weight quantization and distillation: it can be applied alongside them, preserving its inference-time stability gains at the pruning stage without affecting their compression objectives. Favorable results are therefore expected from combining HIES with such methods, and an empirical study of these compositions is left for future work.

References

Joshua Ainslie, James Lee-Thorp, Michiel De Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4895–4901.

Argenis Arriojas, Jacob Adamczyk, Stas Tiomkin, and Rahul V Kulkarni. 2023. Entropy regularized reinforcement learning using large deviation theory. *Physical Review Research*, 5(2):023085.

Amit Ben Artzy and Roy Schwartz. 2024. Attend first, consolidate later: On the importance of attention in different llm layers. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 177–184.

Saleh Ashkboos, Maximilian L Croci, Marcelo Gennari do Nascimento, Torsten Hoefler, and James Hensman. 2024. Slicegpt: Compress large language models by deleting rows and columns. *arXiv preprint arXiv:2401.15024*.

Yonatan Ashlag, Uri Koren, Mirco Mutti, Esther Derman, Pierre-Luc Bacon, and Shie Mannor. 2025. State entropy regularization for robust reinforcement learning. *arXiv preprint arXiv:2506.07085*.

Haoli Bai, Wei Zhang, Lu Hou, Lifeng Shang, Jin Jin, Xin Jiang, Qun Liu, Michael Lyu, and Irwin King. 2021. Binarybert: Pushing the limit of bert quantization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4334–4348.

Anna Bair, Hongxu Yin, Maying Shen, Pavlo Molchanov, and Jose Alvarez. 2023. Adaptive sharpness-aware pruning for robust sparse networks. *arXiv preprint arXiv:2306.14306*.

Homanga Bharadhwaj, Kevin Xie, and Florian Shkurti. 2020. Model-predictive control via cross-entropy and gradient-based optimization. In *Learning for Dynamics and Control*, pages 277–286. PMLR.

Jose Blanchet, Peng Cui, Jiajin Li, and Jiashuo Liu. 2024. Stability evaluation through distributional perturbation analysis. In *Forty-first International Conference on Machine Learning*.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer.

Olivier Bousquet and André Elisseeff. 2002. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526.

Stephen Boyd and Lieven Vandenberghe. 2004. *Convex optimization*. Cambridge university press.

Mengzhao Chen, Wenqi Shao, Peng Xu, Jiahao Wang, Peng Gao, Kaipeng Zhang, and Ping Luo. 2025a. Efficientqat: Efficient quantization-aware training for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10081–10100.

Yingfa Chen, Yutong Wu, Chenyang Song, Zhen Leng Thai, Xingyu Shen, Xu Han, Zhiyuan Liu, and Maosong Sun. 2025b. Cost-optimal grouped-query attention for long-context modeling. *arXiv preprint arXiv:2503.09579*.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Lucas BV De Amorim, George DC Cavalcanti, and Rafael MO Cruz. 2023. The choice of scaling technique matters for classification performance. *Applied Soft Computing*, 133:109924.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, volume 1 (long and short papers)*, pages 4171–4186.

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Angela Fan, Edouard Grave, and Armand Joulin. 2019. Reducing transformer depth on demand with structured dropout. *arXiv preprint arXiv:1909.11556*.

Tamer Ghattas, Michael Hassid, and Roy Schwartz. 2025. On pruning state-space llms. *arXiv preprint arXiv:2502.18886*.

Mitchell Gordon, Kevin Duh, and Nicholas Andrews. 2020. Compressing bert: Studying the effects of weight pruning on transfer learning. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 143–155.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Shuai Han, Wenbo Zhou, Shuai Lü, Sheng Zhu, and Xiaoyu Gong. 2023. Entropy regularization methods for parameter space exploration. *Information Sciences*, 622:476–489.

Song Han, Jeff Pool, John Tran, and William Dally. 2015. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28.

Moritz Hardt, Ben Recht, and Yoram Singer. 2016. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Torsten Hoefler, Dan Alistarh, Tal Ben-Nun, Nikoli Dryden, and Alexandra Peste. 2021. Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. *Journal of Machine Learning Research*, 22(241):1–124.

Jonghyun Hong and Sungyoon Lee. 2025. Variance sensitivity induces attention entropy collapse and instability in transformers. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8371–8389.

Lu Hou, Zhiqi Huang, Lifeng Shang, Xin Jiang, Xiao Chen, and Qun Liu. 2020. Dynabert: Dynamic bert with adaptive width and depth. *Advances in Neural Information Processing Systems*, 33:9782–9793.

Ghadeer Jaradat, Mohammed Tolba, Ghada Alsuhli, Hani Saleh, Mahmoud Al-Qutayri, Thanos Stouraitis, and Baker Mohammad. 2024. Hybrid dynamic pruning: A pathway to efficient transformer inference. *arXiv preprint arXiv:2407.12893*.

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

Young Geun Kim and Carole-Jean Wu. 2020. Autoscale: Energy efficiency optimization for stochastic edge inference using reinforcement learning. In *2020 53rd Annual IEEE/ACM international symposium on microarchitecture (MICRO)*, pages 1082–1096. IEEE.

Alex Krizhevsky, Geoffrey Hinton, and 1 others. 2009. Learning multiple layers of features from tiny images.

Eldar Kurtic, Daniel Campos, Tuan Nguyen, Elias Frantar, Mark Kurtz, Benjamin Fineran, Michael Goin, and Dan Alistarh. 2022. The optimal bert surgeon: Scalable and accurate second-order pruning for large language models. *arXiv preprint arXiv:2203.07259*.

François Lagunas, Ella Charlaix, Victor Sanh, and Alexander M Rush. 2021. Block pruning for faster transformers. *arXiv preprint arXiv:2109.04838*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7871–7880.

Bingbing Li, Zigeng Wang, Shaoyi Huang, Mikhail A Bragin, Ji Li, and Caiwen Ding. 2023a. Towards lossless head pruning through automatic peer distillation for language models. In *IJCAI*, pages 5113–5121.

Yixiao Li, Yifan Yu, Qingru Zhang, Chen Liang, Pengcheng He, Weizhu Chen, and Tuo Zhao. 2023b. Losparse: Structured compression of large language models based on low-rank and sparse approximation. In *International Conference on Machine Learning*, pages 20336–20350. PMLR.

Felipe Tomazelli Lima and Vinicius MA Souza. 2023. A large comparison of normalization methods on time series. *Big Data Research*, 34:100407.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zejian Liu, Fanrong Li, Gang Li, and Jian Cheng. 2021. Ebert: Efficient bert inference with dynamic structured pruning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4814–4823.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. 2023. Llm-pruner: On the structural pruning of large language models. *Advances in neural information processing systems*, 36:21702–21720.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*.

Dmitry Molchanov, Arsenii Ashukha, and Dmitry Vetrov. 2017. Variational dropout sparsifies deep neural networks. In *International conference on machine learning*, pages 2498–2507. PMLR.

Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. 2019. Importance estimation for neural network pruning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11264–11272.

Saurav Muralidharan, Sharath Turuvekere Sreenivas, Raviraj Joshi, Marcin Chochowski, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, Jan Kautz, and Pavlo Molchanov. 2024. Compact language models via pruning and knowledge distillation. *Advances in Neural Information Processing Systems*, 37:41076–41102.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, and 1 others. 2025. Olmo 3. *arXiv preprint arXiv:2512.13961*.

Haojie Pan, Chengyu Wang, Minghui Qiu, Yichang Zhang, Yaliang Li, and Jun Huang. 2021. Meta-kd: A meta knowledge distillation framework for language model compression across domains. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3026–3036.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the association for computational linguistics*, 8:842–866.

Rajarshi Saha, Naomi Sagan, Varun Srivastava, Andrea Goldsmith, and Mert Pilanci. 2024. Compressing large language models using low rank and low precision decomposition. *Advances in Neural Information Processing Systems*, 37:88981–89018.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Shai Shalev-Shwartz and Shai Ben-David. 2014. *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient knowledge distillation for bert model compression. *arXiv preprint arXiv:1908.09355*.

Siqi Sun, Zhe Gan, Yuwei Fang, Yu Cheng, Shuohang Wang, and Jingjing Liu. 2020. Contrastive distillation on intermediate representations for language model compression. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 498–508.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pages 353–355.

Boshi Wang, Xiang Yue, Yu Su, and Huan Sun. 2024. Grokking of implicit reasoning in transformers: A mechanistic journey to the edge of generalization. *Advances in Neural Information Processing Systems*, 37:95238–95265.

Chaoqi Wang, Roger Grosse, Sanja Fidler, and Guodong Zhang. 2019. Eigendamage: Structured pruning in the kronecker-factored eigenbasis. In *International conference on machine learning*, pages 6566–6575. PMLR.

Jiawei Wang, Jiacai Liu, Yuqian Fu, Yingru Li, Xintao Wang, Yuan Lin, Yu Yue, Lin Zhang, Yang Wang, and Ke Wang. 2025. Harnessing uncertainty: Entropy-modulated policy gradients for long-horizon llm agents. *arXiv preprint arXiv:2509.09265*.

Jeffrey TH Wong, Cheng Zhang, Xinye Cao, Pedro Gimenes, George A Constantinides, Wayne Luk, and Yiren Zhao. 2025. A3: an analytical low-rank approximation framework for attention. *arXiv preprint arXiv:2505.12942*.

Mengzhou Xia, Zexuan Zhong, and Danqi Chen. 2022. Structured pruning learns compact and accurate models. *arXiv preprint arXiv:2204.00408*.

Changnan Xiao, Haosen Shi, Jiajun Fan, and Shihong Deng. 2021. An entropy regularization free mechanism for policy-based reinforcement learning. *arXiv preprint arXiv:2106.00707*.

Han Xiao, Kashif Rasul, and Roland Vollgraf. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Dongkuan Xu, Ian EH Yen, Jinxi Zhao, and Zhibin Xiao. 2021. Rethinking network pruning–under the pre-train and fine-tune paradigm. *arXiv preprint arXiv:2104.08682*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Dingqing Yang, Amin Ghasemazar, Xiaowei Ren, Maximilian Golub, Guy Lemieux, and Mieszko Lis. 2020. Procrustes: a dataflow and accelerator for sparse deep neural network training. In *2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pages 711–724. IEEE.

Yifei Yang, Zouying Cao, and Hai Zhao. 2024. Laco: Large language model pruning via layer collapse. *arXiv preprint arXiv:2402.11187*.

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. 2022. Zeroquant: Efficient and affordable post-training quantization for large-scale transformers. *Advances in neural information processing systems*, 35:27168–27183.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33:5824–5836.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Ofir Zafrir, Guy Boudoukh, Peter Izsak, and Moshe Wasserblat. 2019. Q8bert: Quantized 8bit bert. In *2019 Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing-NeurIPS Edition (EMC2-NIPS)*, pages 36–39. IEEE.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua M Susskind. 2023. Stabilizing transformer training by preventing attention entropy collapse. In *International Conference on Machine Learning*, pages 40770–40803. PMLR.

Zixuan Zhou, Xuefei Ning, Ke Hong, Tianyu Fu, Jiaming Xu, Shiyao Li, Yuming Lou, Luning Wang, Zhihang Yuan, Xiuhong Li, and 1 others. 2024. A survey on efficient inference for large language models. *arXiv preprint arXiv:2404.14294*.

Appendix

Table of Contents

A. Related Work p. 14

B. Proofs and Details p. 15
- B.1. Loss-increase Control via Head Importance (Section 4.2.1) p. 15
  - B.1.1. Loss Bound with Operator-Norm Control p. 15
  - B.1.2. Norms and Inner Products p. 15
  - B.1.3. Estimating $\|W^O\|_2$ and Blockwise Norms via Power Iteration p. 16
  - B.1.4. Cross-Entropy Curvature and Propagation to $\mathbf{y}$ p. 16
  - B.1.5. Why the Quadratic Term Is Typically Negligible p. 17
  - B.1.6. Remarks on HIS with Absolute Values p. 17
- B.2. Generalization Gap and Attention Entropy (Section 4.2.2) p. 17
  - B.2.1. Notation and Token Averaging p. 17
  - B.2.2. Neighboring Datasets and Why They Appear p. 18
  - B.2.3. From Attention Perturbation to Output Perturbation p. 18
  - B.2.4. Entropy–Total Variation (TV) Control p. 18
  - B.2.5. Stability and Generalization p. 19
  - B.2.6. Constants and Practical Remarks p. 19
- B.3. Risk Upper Bound and HIES Minimization (Section 4.2.3) p. 19
- B.4. Orthogonality and Complementarity (Section 4.2.4) p. 19

C. Experimental Setup p. 21
- C.1. Experimental Setup for Motivation Study p. 21
- C.2. Model and Datatset p. 21
- C.3. Computing Resources p. 21
- C.4. Dataset Statistics p. 21
  - C.4.1. Natural Language Understanding Task p. 21
  - C.4.2. Image Classification Task p. 21
  - C.4.3. Reasoning Task p. 21
  - C.4.4. Multi-modal Vision-Language Task p. 21

D. Additional Experimental Results p. 25
- D.1. Orthogonality Analysis p. 25
- D.2. Seed Sensitivity and Statistical Significance p. 26
- D.3. Effect of Calibration Dataset Size p. 26
- D.4. Heatmap of Importance Scores and Pruning Results p. 26
- D.5. 3D Analysis of Attention Head Importance Scores p. 26
  - D.5.1. Difference in Pruning Patterns p. 26
  - D.5.2. Performance Inversion Across Sparsity Regimes p. 26
- D.6. Experimental Results on Large-scale Reasoning Tasks p. 26
- D.7. Experimental Results on Downstream Tasks p. 26
- D.8. Sensitivity Analysis - Ablation on α p. 27
- D.9. Evaluation with a Single Fixed α Across Model Families p. 28
- D.10. Combining Attention Entropy with Other Importance Signals p. 28

E. Efficiency Analysis p. 28
- E.1. Computational Efficiency and FLOPs Reduction p. 28
- E.2. Runtime Overhead and Implementation Details p. 29

A Related Work

Model Compression. Language models (Devlin et al., 2019; Liu et al., 2019; Lewis et al., 2020) have rapidly advanced in scale and capability, intensifying the demand to reduce their parameter sizes and inference latency (Molchanov et al., 2017; Gordon et al., 2020). To compress large language models efficiently, various approaches have been explored, including pruning (Liu et al., 2021; Kurtic et al., 2022; Xu et al., 2021), knowledge distillation (Sun et al., 2019, 2020; Pan et al., 2021), quantization (Bai et al., 2021; Yao et al., 2022; Zafrir et al., 2019; Chen et al., 2025a), and other techniques, like low-rank approximation methods (Saha et al., 2024; Li et al., 2023b; Wong et al., 2025) or weight space decomposition methods such as (Ashkboos et al., 2024).

Structured Pruning. Among these, structured pruning—which removes entire architectural components rather than individual weights— can be performed at various granularities, such as whole layers (Fan et al., 2019), multi-head attention modules (Michel et al., 2019; Voita et al., 2019), or feed-forward networks (Lagunas et al., 2021; Hou et al., 2020). Recently, attention head pruning has gained particular traction in the context of large language models. It directly reduces attention FLOPs and KV-cache size by lowering the number of active heads while preserving the layer topology, thereby simplifying checkpoint compatibility and serving integration. Consequently, several large-scale studies adopt head-level pruning as a practical axis in LLM compression pipelines (Ma et al., 2023; Jaradat et al., 2024; Muralidharan et al., 2024). Ma et al. (2023) propose a unified framework that integrates structured head pruning into the training pipeline of large language models, achieving substantial sparsity without accuracy degradation.

Entropy in Reinforcement Learning and Transformers. Entropy has long been employed in reinforcement learning (RL) as a means of encouraging exploration and preserving policy diversity. By introducing an entropy term into the policy objective, RL methods prevent premature convergence to deterministic strategies and mitigate policy collapse (Bharadhwaj et al., 2020; Xiao et al., 2021; Wang et al., 2025). Extensions of this idea include parameter-space entropy regularization to explicitly control diversity (Han et al., 2023), large-deviation interpretations of entropy-regularized RL (Arriojas et al., 2023), and state-distribution entropy regularization for improved robustness and generalization (Ashlag et al., 2025).

In Transformer architectures, attention entropy (AE) has been introduced to quantify the concentration of each head's focus across input tokens. Zhai et al. (2023) report that persistently low AE—an "entropy collapse" state—correlates strongly with training instabilities such as oscillations in the loss surface and divergence across model scales. Their theoretical analysis ties AE to the spectral norm of attention logits, and they propose σReparam to prevent collapse by enforcing a lower bound on entropy. More recent work identifies variance sensitivity in the softmax transformation as another source of entropy collapse (Hong and Lee, 2025). Together, these findings highlight attention entropy as a critical factor for stability, motivating its integration as a complementary signal in pruning frameworks.

B Proofs and Details

Overview. This appendix collects the formal analyses that underpin Section 4.2 and fixes notation and technical conventions used throughout. We first develop loss-increase control for head pruning via gradient-based 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 B.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 B.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 B.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 B.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.

B.1 Loss-increase Control via Head Importance (Section 4.2.1)

B.1.1 Loss Bound with Operator-Norm Control

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

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

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 $\delta\mathbf{y} = \mathrm{Concat}(\delta A_1, \ldots, \delta A_{|\mathrm{H}|})$, $\|\delta\mathbf{y}\|_F^2 = \sum_h \|\delta A_h\|_F^2$ 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

$$\Delta\mathcal{L} \;\le\; \sum_{h=1}^{|\mathrm{H}|} (1 - m_h)\, \mathrm{HIS}_h \;+\; \frac{\beta_y}{2} \sum_{h=1}^{|\mathrm{H}|} (1 - m_h)\, \|A_h\|_F^2. \tag{9}$$

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

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

B.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 $\frac{1}{Bn}$).

B.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 $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$, 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

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

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.

B.1.4 Cross-Entropy Curvature and Propagation to y

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

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

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

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

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

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, $\mathrm{Var}_p(v) \leq \frac{(\max_i v_i - \min_i v_i)^2}{4} \leq \frac{1}{2}$ 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 $\mathbf{z} = \mathbf{y}W^O$,

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

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

B.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 B.1.2. Assume there exists $g > 0$ such that, for all heads under consideration,

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

e.g., define

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

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

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

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

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

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

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 B.1.1.

B.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 B.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.

B.2 Generalization Gap and Attention Entropy (Section 4.2.2)

B.2.1 Notation and Token Averaging

For head h and query token $t \in \{1,\ldots,n(x)\}$, 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 B.2.1). Here $(\mathcal{S},\mathcal{S}')$ are *neighboring* datasets that differ in one example.

B.2.2 Neighboring Datasets and Why They Appear

We call two datasets $S = (z_1, \ldots, z_N)$ and $S' = (z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_N)$ *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.

B.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 $\|V_h\|_{\infty\to 2} \le M$,

$$\|\Delta_h(t)\|_2 \;\le\; \|\boldsymbol{\alpha}_t^{(h)} - \boldsymbol{\alpha}_t'^{(h)}\|_1 \cdot \|V_h\|_{\infty\to 2} \;\le\; M\,\|\boldsymbol{\alpha}_t^{(h)} - \boldsymbol{\alpha}_t'^{(h)}\|_1. \tag{10}$$

Averaging over tokens and applying the mask m_h,

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

B.2.4 Entropy–Total Variation (TV) Control

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

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

Applying Lemma B.1 to (10) token-wise and averaging,

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

Therefore,

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

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

$$\|\Delta(x)\|_2 \;\le\; \underbrace{\sqrt{8}\,M\,\sqrt{|\mathrm{H}|\rho\,\log n(x)}}_{=:\,C_{\mathrm{AE}}(x)}\cdot\sqrt{\sum_{h=1}^{|\mathrm{H}|}(1-m_h)\,\overline{\mathrm{AD}}_h(x)}. \tag{11}$$

B.2.5 Stability and Generalization

Let $\gamma := L_\ell \, \mathbb{E}_{\mathcal{S}}[\|\Delta(x)\|_2]$. By on-average replace-one stability (Bousquet and Elisseeff, 2002),

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

Using (11) and Jensen for $\sqrt{\cdot}$,

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

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

B.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.

B.3 Risk Upper Bound and HIES Minimization (Section 4.2.3)

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 $\mathrm{HIES}_j > \mathrm{HIES}_i$. Consider the mask $\tilde{m}$ that swaps i and j (retain j, prune i); the constraint in (7) is preserved. The objective in (6) changes by

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

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 $|\mathrm{H}| - k$ smallest HIES is optimal. □

B.4 Orthogonality and Complementarity (Section 4.2.4)

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 $\langle \mathbb{E}_{\mathcal{S}}[u_h],\ \mathbb{E}_{\mathcal{S}}[v_h]\rangle = 0$.

C Experimental Setup

C.1 Experimental Setup for Motivation Study

We analyze accuracy degradation and head behaviors under HIS-based pruning. In our diagnostic study, we analyze the phenomena of pruning by HIS on BERT, focusing on detailed attention head behaviors during inference. Following prior work analyzing BERT's attention geometry and mechanisms (Clark et al., 2019; Rogers et al., 2020; Wang et al., 2024), we further explore attention head pruning dynamics.

C.2 Model and Dataset

Model. We use publicly available $\text{BERT}_{\text{base}}$ checkpoints that have been fine-tuned by prior work (Devlin et al., 2019), and $\text{LLaMA-2}_{\text{7B}}$ checkpoint from Hugging Face (Touvron et al., 2023). To examine the generalizability to attention variants and tasks, we further employ $\text{ViT}_{\text{Large}}$ (Dosovitskiy, 2020), $\text{LLaVA-1.5}_{\text{7B}}$ (Liu et al., 2023), Qwen2.5-3B / 7B (Qwen et al., 2025), Qwen3-1.7B / 8B (Yang et al., 2025), $\text{LLaMA-3.1}_{\text{8B}}$ (Grattafiori et al., 2024), DeepSeek-R1-Distill-LLaMA-8B (Guo et al., 2025), and $\text{OLMo-3}_{\text{7B}}$ (Olmo et al., 2025).

Datasets. We evaluate on various widely-adopted benchmarks: GLUE (Wang et al., 2018), HellaSwag (Zellers et al., 2019), Winogrande (Sakaguchi et al., 2021), the AI2 Reasoning Challenge—ARC-e/ARC-c (Clark et al., 2018), OBQA (Mihaylov et al., 2018), ImageNet1k (Deng et al., 2009), CIFAR-100 (Krizhevsky et al., 2009), Food-101 (Bossard et al., 2014), Fashion MNIST (Xiao et al., 2017), MMLU (Hendrycks et al., 2020), GSM8K (Cobbe et al., 2021), VizWiz-VQA (Gurari et al., 2018), and MM-Vet (Yu et al., 2023).

C.3 Computing Resources

Our experimental setup leverages two RTX 4090 GPUs with 24GB memory for NLU tasks using BERT and for image classification tasks using ViT. Experiments involving LLMs such as LLaMA and multi-modal VLMs such as LLaVA were conducted on H100 GPU with 80GB memory. For the MM-Vet benchmark, we evaluated model responses using the OpenAI API to handle open-ended answer scoring.

C.4 Dataset Statistics

C.4.1 Natural Language Understanding Task

We present the dataset statistics of GLUE (Wang et al., 2018) in Table 5.

C.4.2 Image Classification Task

Table 5 lists dataset statistics for the image classification task in the Computer Vision (CV) domain.

C.4.3 Reasoning Task

To evaluate the effectiveness of our pruning method on reasoning tasks, we use two benchmark datasets: **GSM8K** (Cobbe et al., 2021) and **MMLU** (Hendrycks et al., 2020).

GSM8K: is a dataset of 8.5K high quality, linguistically diverse grade school math word problems. GSM8K supports the task of question answering on basic mathematical problems requiring multi-step reasoning.

MMLU: is a massive multitask benchmark consisting of multiple-choice questions spanning diverse domains, including the humanities, social sciences, and hard sciences. The benchmark covers 57 tasks such as elementary mathematics, U.S. history, computer science, and law, and requires models to possess broad world knowledge and strong problem-solving abilities to achieve high performance.

By using both GSM8K and MMLU, we evaluate our pruning methods from complementary perspectives: GSM8K assesses mathematical reasoning under multi-step computation, while MMLU measures the preservation of broad world knowledge and multi-domain problem-solving ability.

C.4.4 Multi-modal Vision-Language Task

To evaluate the effectiveness of our pruning method on multi-modal vision-language models (VLMs), we use two benchmark datasets: **VizWiz-VQA** (Gurari et al., 2018) and **MM-Vet** (Yu et al., 2023). The evaluation was conducted using the $\text{LLaVA1.5}_{\text{7B}}$ model.

VizWiz-VQA: is designed for Visual Question Answering (VQA) in the context of assisting people who are blind. Each visual question originates from a real-world setting where blind users captured images and recorded spoken questions, and is accompanied by ten crowdsourced answers. The dataset poses two evaluation tasks: predicting the correct answer given an image and question, and detecting whether a question cannot be answered.

MM-Vet: is a benchmark for evaluating large multimodal models on complex tasks requiring

Table 4: Summary of model parameters and architectures.

Model	Parameters	# Layers	# Attention Heads	Architecture / Key Details
$BERT_{base}$	110M	12	12	Transformer encoder; pre-trained on masked language modeling and next sentence prediction tasks.
$LLaMA\text{-}2_{7B}$	7B	32	32	Transformer decoder-only; trained on large-scale text corpora for general-purpose language modeling.
ViT_{Large}	307M	24	16	Vision Transformer; patch-based image tokenization (16×16), pre-trained on ImageNet for image classification tasks.
$LLaVA\text{-}1.5_{7B}$	7B	32	32	Multi-modal LLaMA variant integrating a visual encoder; capable of joint image-text understanding and generation.

Table 5: Summary of the NLU and CV benchmarks.

NLU Benchmark						
Dataset	# Train	# Valid	# Test	# Label	Task	Evaluation Metric
Single-Sentence Classification (GLUE)						
CoLA	8,551	521	522	2	Acceptability	Matthews corr
SST-2	66,349	1,000	872	2	Sentiment	Accuracy
Pairwise Text Classification (GLUE)						
MNLI	392,702	9,832	9,815	3	NLI	Accuracy
RTE	2,490	138	139	2	NLI	Accuracy
QQP	362,846	1,000	40,431	2	Paraphrase	Accuracy
MRPC	3,668	204	204	2	Paraphrase	F1 score
QNLI	103,743	1,000	5,463	2	QA/NLI	Accuracy
Regression (GLUE)						
STS-B	5,749	750	750	1	Similarity	Pearson corr
CV Benchmark						
Dataset	# Train	# Valid	# Test	# Label	Task	Evaluation Metric
ImageNet1k	1,281,167	50,000	100,000	1,000	Classification	Accuracy
CIFAR-100	45,000	5,000	10,000	100	Classification	Accuracy
Fashion MNIST	54,000	6,000	10,000	10	Classification	Accuracy
Oxford Flowers	1,020	1,020	6,150	102	Classification	Accuracy

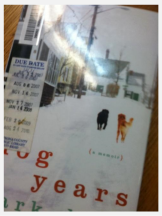
multiple vision-language capabilities. It defines six core VL skills and their combinations, using an LLM-based evaluator for unified scoring across diverse question and answer formats.

By using both VizWiz-VQA and MM-Vet, we evaluate our pruning method across real-world visual questions, multimodal reasoning, and diverse answer styles. Note that our evaluation is conducted on a subset of the datasets, and we provide example entries from both benchmarks for illustration.

Q: Can you tell me what this medicine is please?

Answers: [
"no", "unanswerable",
"night time",
"unanswerable",
"night time",
"night time cold medicine",
"night time",
"night time",
"night time",
"night time medicine"
]

Category: Other



Q: What is the title of this book?

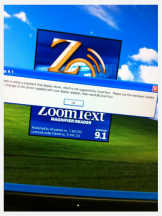
Answers: [
"dog years",
"dog years: memoir",
"dog years",
"dog years",
"dog years",
"dog years",
"dog years",
"dog years",
"dog years",
"dog years"
]

Category: Other

Q: Which one is the blue one?

Answers: [
"right",
"right",
"right",
"omni sorb",
"right",
"1 on right",
"on right",
"right",
"right 1",
"right 1"
]

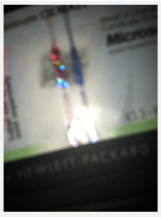
Category: Other



Q: What does the arrow say?

Answers: [
"unanswerable",
"unanswerable",
"zoomtext",
"unanswerable",
"unanswerable",
"unanswerable",
"unanswerable",
"unanswerable",
"zoomtext"
]

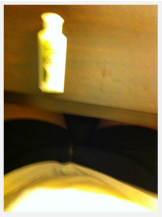
Category: Unanswerable



Q: Okay I think I got it this time. Can you read the model and serial number for me? Thanks.

Answers: [
"unanswerable",
"unanswerable",
"unanswerable",
"unanswerable",
"unanswerable",
"no",
"unanswerable",
"unanswerable",
"unanswerable",
"unanswerable"
]

Category: Unanswerable



Q: Is this shampoo or conditioner?

Answers: [
"unanswerable",
"unanswerable",
"unanswerable",
"unanswerable",
"unanswerable",
"unanswerable",
"unanswerable",
"unanswerable",
"unanswerable",
"unanswerable"
]

Category: Unanswerable

Q: Can you describe for me what's going on outside the window? Just curious.

Answers: [
"store front",
"unanswerable",
"nothing going outside window just building windows sign to right",
"yes",
"nothing",
"nothing",
"nothing",
"nothing",
"sun shining on building on corner trees next to",
"building"
]

Category: Other

Q: What about the signery.

Answers: [
"unanswerable",
"unanswerable",
"political",
"man",
"unanswerable",
"no clue",
"unanswerable",
"unanswerable",
"unanswerable",
"blue man in on building"
]

Category: Unanswerable

Q: This piece of mail for and where is it from?

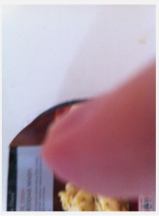
Answers: [
"unanswerable",
"unanswerable",
"unanswerable",
"unanswerable",
"unanswerable",
"no",
"unanswerable",
"unanswerable",
"cannot see full size envelope on this image",
"unanswerable"
]

Category: Unanswerable

Q: Which one of these items is the children's dictionary? Is it the one on the right, or the one on the left?

Answers: [
"left",
"left",
"left",
"left",
"left",
"right",
"left",
"left",
"left",
"left"
]

Category: Other



Q: What is this?

Answers: [
"finger",
"unanswerable",
"unanswerable",
"person",
"unanswerable",
"unanswerable",
"unanswerable",
"instructions",
"thumb",
"finger"
]

Category: Other

Q: Can you see if this is 32 bit or 64 bit? Thanks.

Answers: [
"dog years",
"dog years: memoir",
"dog years",
"dog years",
"dog years",
"dog years",
"dog years",
"dog years",
"dog years",
"dog years"
]

Category: Unanswerable

Figure 5: Examples from VizWiz-VQA showing visual questions asked by blind users and the corresponding answers from crowd workers. The examples include both questions that can be answered from the image and questions that cannot.

Q: How many gallons of supreme gasoline can I get with $50?

GT: 13.6 <OR> 13.7

Required capabilities: OCR, math

Q: On the right desk, what is to the left of the laptop?

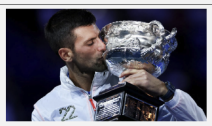
GT: table lamp <OR> desk lamp

Required capabilities: Recognition, spatial awareness

Q: What occasions would someone use this meme?

GT: This meme, commonly known as "Screaming Panda," is typically used to express shock, surprise, or fear. It could be used in response to a startling or unexpected event, or to convey a sense of panic or alarm. Some possible occasions where someone might use this meme include:- Reacting to a jump scare in a horror movie- Responding to a surprising plot twist in a TV show or book- Expressing shock at a news headline or current event- Conveying fear or anxiety about an upcoming deadline or exam- Showing surprise at an unexpected outcome in a sports game or other competition.

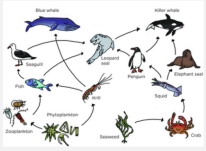
Required capabilities: Recognition, knowledge, language generation



Q: In which country was this photo taken?

GT: Australia

Required capabilities: Recognition, knowledge



Q: Which are producers in this food web?
GT: Phytoplankton <AND> Seaweed
Required capabilities: OCR, knowledge, spatial awareness

Underground Railway Systems

City	Date opened	Kilometres of route	Passengers per year (in millions)
London	1863	394	775
Paris	1900	199	1191
Tokyo	1927	155	1927
Washington DC	1976	126	144
Kyoto	1981	11	45
Los Angeles	2001	28	50

Q: The table below gives information about the underground railway systems in six cities. Summarize the information by selecting and reporting the main features and make comparisons where relevant. You should write at least 150 words.

GT: The table shows data about the underground rail networks in six major cities. The table compares the six networks in terms of their age, size and the number of people who use them each year. It is clear that the three oldest underground systems are larger and serve significantly more passengers than the newer systems. The London underground is the oldest system, having opened in 1863. It is also the largest system, with 394 kilometers of route. The second largest system, in Paris, is only about half the size of the London underground, with 199 kilometers of route. However, it serves more people per year. While only third in terms of size, the Tokyo system is easily the most used, with 1927 million passengers per year. Of the three newer networks, the Washington DC underground is the most extensive, with 126 kilometers of route, compared to only 11 kilometers and 28 kilometers for the Kyoto and Los Angeles systems. The Los Angeles network is the newest, having opened in 2001, while the Kyoto network is the smallest and serves only 45million passengers per year.

Required capabilities: OCR, language generation, spatial awareness

Q: How many tomatoes are there?
GT: 5
Required capabilities: Recognition

Q: The graph below shows the long-term international migration, UK, 1999-2008.Summarize the information by selecting and reporting the main features and make comparisons where relevant. You should write at least 150 words.

GT: The chart gives information about UK immigration, emigration and net migration between 1999 and 2008.Both immigration and emigration rates rose over the period shown, but the figures for immigration were significantly higher. Net migration peaked in 2004 and 2007.In 1999, over 450,000 people came to live in the UK, while the number of people who emigrated stood at just under 300,000. The figure for net migration was around160,000, and it remained at a similar level until 2003. From 1999 to 2004, the immigration rate rose by nearly 150,000 people, but there was a much smaller rise in emigration. Net migration peaked at almost 250,000 people in 2004. After 2004, the rate of immigration remained high, but the number of people emigrating fluctuated. Emigration fell suddenly in 2007, before peaking at about 420,000 people in 2008. As a result, the net migration figure rose to around 240,000in 2007, but fell back to around 160,000 in 2008.

Required capabilities: Recognition, OCR, language generation, spatial awareness

Figure 6: Eight example queries from the MM-Vet benchmark, each requiring different integrations of core vision–language capabilities to solve complicated multimodal tasks.

Table 6: Orthogonality analysis using $\text{BERT}_{\text{base}}$ on the GLUE benchmark.

Task	E[ũ]	E[ṽ]	Cov(ũ,ṽ)	E[⟨ũ,ṽ⟩]	Correlation
COLA	6.94E-03	4.08E-01	-1.40E-05	2.82E-03	-1.54E-02
SST2	6.94E-03	4.42E-01	-3.60E-05	3.03E-03	-3.41E-02
MRPC	6.94E-03	5.27E-01	2.41E-04	3.90E-03	2.32E-01
STSB	6.94E-03	4.30E-01	-1.96E-04	2.79E-03	-1.65E-01
QQP	6.94E-03	5.05E-01	1.56E-04	3.67E-03	1.95E-01
MNLI	6.94E-03	4.89E-01	1.73E-04	3.57E-03	2.27E-01
QNLI	6.94E-03	4.62E-01	1.53E-04	3.36E-03	2.23E-01
RTE	6.94E-03	5.05E-01	2.15E-04	3.72E-03	2.88E-01

Table 7: Orthogonality analysis using $\text{LLaMA-2}_{\text{7B}}$ on 5 tasks.

Task	E[ũ]	E[ṽ]	Cov(ũ,ṽ)	E[⟨ũ,ṽ⟩]	Correlation
HellaSwag	9.77E-04	7.56E-01	-1.47E-04	5.92E-04	-1.44E-01
Winogrande	9.77E-04	7.66E-01	-1.85E-04	5.63E-04	-1.66E-01
ARC-e	9.77E-04	7.76E-01	-1.42E-04	6.16E-04	-1.44E-01
ARC-c	9.77E-04	7.71E-01	-1.27E-04	6.26E-04	-1.33E-01
OBQA	9.77E-04	8.01E-01	-1.29E-04	6.53E-04	-1.61E-01

D Additional Experimental Results

D.1 Orthogonality Analysis

We provide an empirical sanity check supporting the Lemma 4.3. Experiments use TinyBERT on SST-2 (`Vishnou/TinyBERT_SST2`). For each head we compute layerwise-normalized HIS and 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:

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

Consequently,

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

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 further extend the orthogonality analysis across models, with results reported in Table 6 and Table 7.

D.2 Seed Sensitivity and Statistical Significance

We evaluate the seed sensitivity of HIS and HIES, with results reported in Table 8 and Table 9 .

D.3 Effect of Calibration Dataset Size

We analyze the effect of calibration dataset size on the performance and stability of HIES. Tables 10 and 11 report accuracy and standard deviation across calibration sizes ranging from 1 to 1024 samples for $\text{BERT}_{\text{base}}$ and $\text{LLaMA-2}_{\text{7B}}$, respectively. Across both model families, HIES exhibits stable performance once a modest number of calibration samples is used. In particular, increasing the calibration set size beyond small values yields diminishing returns in accuracy, while the variance across random seeds remains low. These results indicate that HIES does not require large calibration datasets to obtain reliable head-level statistics, supporting its practical applicability in low-cost calibration settings.

D.4 Heatmap of Importance Scores and Pruning Results

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.

D.5 3D Analysis of Attention Head Importance Scores

D.5.1 Difference in Pruning Patterns

As shown in Figure 9, Pruning heatmaps 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.

D.5.2 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 HIS is 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 $\Rightarrow$ HIS outperforms HIES.
- **Pruning $\geq$ 30%:** Specialization regime $\Rightarrow$ 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.

D.6 Experimental Results on Large-scale Reasoning Tasks

Figure 10 reports reasoning performance on GSM8K and MMLU with $\text{LLaMA-2}_{\text{7B}}$, comparing HIES against HIS-based pruning. Across both benchmarks, HIES consistently outperforms HIS across pruning regimes, achieving an average accuracy improvement of 14.67% and demonstrating improved robustness on reasoning tasks.

D.7 Experimental Results on Downstream Tasks

Figure 11 reports downstream evaluations of HIES versus the HIS baseline on CIFAR-100, Food-101, and Fashion-MNIST. Across all three benchmarks, HIES consistently sustains higher accuracy under aggressive pruning, whereas HIS exhibits rapid degradation once the pruning ratio exceeds 20%.

Table 8: Accuracy and standard deviation in accuracy across 5 random seeds for $\text{BERT}_{\text{base}}$ on GLUE benchamrk.

	Pruning Ratio (%)	HIS		HIES (ours)	
		Accuracy (%)	Standard Deviation	Accuracy (%)	Standard Deviation
CoLA	10	74.18	1.60E-02	75.85	1.21E-02
	20	74.18	5.44E-04	73.04	1.09E-02
	30	72.70	6.91E-03	75.17	2.82E-03
	40	70.28	5.52E-03	70.20	3.68E-02
	50	60.37	2.18E-02	71.17	5.29E-03
MNLI	10	84.13	0.00E+00	84.06	1.64E-01
	20	82.93	9.13E-02	82.87	0.00E+00
	30	82.40	0.00E+00	82.29	3.58E-01
	40	81.46	2.39E-01	81.39	1.98E-01
	50	79.47	4.34E-01	78.87	3.27E-01
MRPC	10	90.29	4.58E-03	90.97	8.04E-04
	20	89.69	4.50E-03	89.17	3.55E-03
	30	88.79	2.36E-03	89.52	3.69E-03
	40	88.02	7.41E-03	87.12	9.06E-03
	50	86.85	1.65E-02	87.19	1.66E-02
QNLI	10	90.55	8.69E-02	90.51	1.34E-01
	20	90.23	3.48E-01	90.05	2.73E-01
	30	88.33	7.47E-01	88.15	7.43E-01
	40	84.75	1.45E+00	85.02	8.90E-01
	50	80.81	3.51E+00	82.83	1.74E+00
QQP	10	90.68	2.33E-02	91.26	8.00E-03
	20	90.56	1.52E-01	90.56	7.31E-02
	30	89.84	3.67E-01	89.72	3.45E-01
	40	88.07	1.84E-01	87.96	2.54E-01
	50	85.33	9.93E-01	85.90	6.97E-01
RTE	10	72.42	2.89E-01	72.13	2.70E-01
	20	71.91	5.78E-01	70.76	9.69E-01
	30	71.26	1.18E+00	71.34	1.47E+00
	40	67.94	2.10E+00	66.79	1.92E+00
	50	65.99	1.36E+00	65.42	1.56E+00
SST-2	10	91.86	1.62E-01	92.64	4.59E-02
	20	91.77	2.22E-01	92.11	1.69E-01
	30	91.19	3.03E-01	92.20	1.92E-01
	40	88.67	1.62E+00	90.96	7.16E-01
	50	86.31	1.35E+00	90.55	1.24E+00
STS-B	10	87.85	1.88E-03	87.95	3.37E-03
	20	87.33	3.52E-03	87.35	2.06E-04
	30	86.82	2.15E-03	86.76	3.25E-03
	40	85.89	4.49E-03	85.87	4.41E-03
	50	83.38	8.95E-03	82.27	1.37E-02
Avg.	10	85.25	7.30E-02	85.67	7.98E-02
	20	84.82	1.75E-01	84.49	1.87E-01
	30	83.92	3.26E-01	84.39	3.90E-01
	40	81.89	7.00E-01	81.91	5.04E-01
	50	78.56	9.62E-01	80.53	6.99E-01

On CIFAR-100, HIS collapses beyond moderate sparsity, while HIES exhibits slower degradation and retains substantially higher accuracy relative to HIS even at 40–50%. Food-101 reveals a similar trend, with HIES delivering substantial and consistent gains over HIS across all pruning levels. On Fashion-MNIST, HIS undergoes steep drops after 20% pruning, in contrast to the stable performance of HIES up to 50%. These results demonstrate that HIES reliably mitigates sharp-drop phenomena and delivers robust, stable improvements over HIS across heterogeneous downstream tasks.

D.8 Sensitivity Analysis - Ablation on α

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

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

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 12.

Table 9: Accuracy and standard deviation in accuracy across 5 random seeds for LLaMA-2_{7B} on HellaSwag, Winogrande, and ARC-c.

	Pruning Ratio (%)	HIS		HIES (ours)	
		Accuracy (%)	Standard Deviation	Accuracy (%)	Standard Deviation
HellaSwag	10	53.63	5.36E-01	53.98 0.66%↑	5.40E-01
	20	51.70	5.17E-01	52.36 1.28%↑	5.24E-01
	30	49.00	4.90E-01	49.84 1.71%↑	4.98E-01
	40	45.27	4.53E-01	46.27 2.20%↑	4.63E-01
	50	37.01	3.70E-01	36.13 2.37%↓	3.61E-01
Winogrande	10	66.38	6.64E-01	67.01 0.95%↑	6.70E-01
	20	64.07	6.41E-01	66.27 3.42%↑	6.63E-01
	30	63.03	6.30E-01	64.67 2.60%↑	6.47E-01
	40	60.16	6.02E-01	61.06 1.50%↑	6.11E-01
	50	56.16	5.62E-01	55.83 0.59%↓	5.58E-01
ARC-c	10	33.15	3.71E-03	35.05 5.73%↑	7.03E-03
	20	33.15	5.57E-03	36.27 9.41%↑	0.00E+00
	30	32.34	1.86E-03	35.46 9.64%↑	5.67E-03
	40	29.69	1.86E-03	34.24 15.30%↑	1.15E-02
	50	25.29	7.43E-03	28.95 14.48%↑	4.22E-02
Avg.	10	51.05	5.70E-03	52.01 1.88%↑	5.06E-03
	20	49.64	5.12E-03	51.63 4.01%↑	2.72E-03
	30	48.12	5.30E-03	49.99 3.88%↑	6.11E-03
	40	45.04	5.49E-03	47.19 4.76%↑	8.82E-03
	50	39.49	9.20E-03	40.30 2.07%↑	2.17E-02

D.9 Evaluation with a Single Fixed α Across Model Families

The main results in Section 5.1 use a per-task tuned α. To verify that HIES's gains do not hinge on this tuning, we additionally evaluate it with a *single* α per model, fixed across all tasks and pruning ratios, on seven recent decoder-only models spanning Grouped-Query Attention (Ainslie et al., 2023) and standard MHA, with parameter sizes from 1.7B to 8B: Qwen2.5-3B / 7B (Qwen et al., 2025), Qwen3-1.7B / 8B (Yang et al., 2025), LLaMA-3.1_{8B} (Grattafiori et al., 2024), DeepSeek-R1-Distill-LLaMA-8B (Guo et al., 2025), and OLMo-3_{7B} (Olmo et al., 2025). In GQA, each query head still produces its own attention distribution over keys, so $\boldsymbol{\alpha}_t^{(h)}$ and AE_h remain well-defined per query head and HIES applies without modification. Across all seven models and three pruning ratios, HIES outperforms HIS on the macro-averaged score, with the gap widening at more aggressive pruning—consistent with the specialization-regime analysis in Appendix D.5.2.

D.10 Combining Attention Entropy with Other Importance Signals

To examine the applicability of AE beyond gradient-based scores, we combine it with a different importance signal based on the L2 norm of attention outputs. This experiment assesses the generality of AE and verifies that it can serve as a complementary stabilization term under different importance signals.

Figure 13 reports the impact of combining AE with the L2-norm–based importance score on LLaMA-2_{7B} across five benchmarks: HellaSwag, Winogrande, ARC-e, ARC-c, and OBQA. Across all benchmarks, incorporating AE consistently mitigates performance degradation under aggressive pruning, indicating that the stabilizing effect of AE is not restricted to a particular importance formulation.

E Efficiency Analysis

E.1 Computational Efficiency and FLOPs Reduction

As the pruning ratio increases, the total FLOPs decrease approximately linearly: removing 10% of attention heads yields an ≈4% reduction in FLOPs, while pruning 50% of heads achieves an ≈20% reduction. Under HIS-only pruning, model accuracy on TinyBERT with SST-2 sharply degrades beyond a 42% pruning ratio, at which point only an ≈16% FLOPs reduction can be attained without critical performance loss. In contrast, HIES-based pruning

Table 10: Accuracy and standard deviation in accuracy for different calibration dataset sizes (1, 16, 32, 64, 128, 512, 1024) for $BERT_{base}$ on GLUE Benchmarks. Note that we use a default size of 32 in the main results.

	Pruning Ratio (%)	HIS		HIES (ours)	
		Accuracy (%)	Standard Deviation	Accuracy (%)	Standard Deviation
CoLA	10	74.46	1.33E-02	75.71	1.34E-02
	30	74.09	8.40E-03	75.10	3.04E-03
	50	62.01	3.37E-02	71.04	1.13E-02
MNLI	10	84.13	0.00E+00	84.13	0.00E+00
	30	82.40	0.00E+00	82.35	3.48E-01
	50	78.93	5.72E-01	78.65	1.98E-01
MRPC	10	90.52	2.82E-03	91.05	2.08E-03
	30	89.12	5.97E-03	89.61	2.94E-03
	50	86.45	1.55E-02	86.92	1.44E-02
QNLI	10	90.51	9.57E-02	90.54	1.25E-01
	30	88.25	5.97E-01	88.01	5.53E-01
	50	80.50	2.96E+00	81.97	1.39E+00
QQP	10	90.74	7.25E-02	91.24	4.16E-02
	30	90.12	2.31E-01	90.01	2.37E-01
	50	84.75	6.91E-01	85.60	4.29E-01
RTE	10	72.36	4.09E-01	72.25	3.25E-01
	30	71.17	1.13E+00	70.81	1.13E+00
	50	65.45	1.14E+00	65.50	2.13E+00
SST-2	10	91.79	1.30E-01	92.56	1.68E-01
	30	90.78	4.53E-01	92.12	2.06E-01
	50	85.76	2.30E+00	90.33	1.16E+00
STS-B	10	87.98	7.08E-04	87.96	2.88E-03
	30	86.87	2.03E-03	86.75	2.64E-03
	50	83.53	6.39E-03	82.80	1.37E-02
Avg.	10	85.31	9.06E-02	85.68	8.48E-02
	30	84.10	3.04E-01	84.34	3.10E-01
	50	78.42	9.63E-01	80.35	6.68E-01

maintains at least 80% validation accuracy up to a 60% pruning ratio, corresponding to an ≈23% FLOPs reduction relative to the original model. Extending the feasible pruning regime from 42% to 60% therefore delivers an additional ≈7 percentage-point reduction in FLOPs, demonstrating that HIES enables substantially greater computational efficiency without compromising task performance.

E.2 Runtime Overhead and Implementation Details

The above efficiency gains are achieved with only a small computational overhead. HIES is computed using lightweight forward hooks attached to the attention modules. Specifically, HIS reuses gradients produced during standard backpropagation, while AE is obtained directly from attention probabilities in the forward pass. Both signals are collected via non-intrusive forward hooks without modifying model internals or introducing additional forward or backward passes. As a result, the runtime overhead is minimal—typically on the order of a few seconds— with negligible additional memory cost beyond a small buffer for storing per-head statistics. Despite this low overhead, HIES consistently outperforms baselines that rely on full fine-tuning, achieving superior efficiency in both computation and accuracy.

Table 11: Accuracy and standard deviation in accuracy for different calibration dataset sizes (1, 16, 32, 64, 128, 512, 1024) for LLaMA-2_{7B} on HellaSwag, Winogrande, and ARC-c. Note that we use a default size of 32 in the main results.

	Pruning Ratio (%)	HIS		HIES (ours)	
		Accuracy (%)	Standard Deviation	Accuracy (%)	Standard Deviation
HellaSwag	10	53.54	1.57E-03	54.09 1.02%↑	1.99E-03
	20	51.44	3.51E-03	52.37 1.82%↑	1.93E-03
	30	49.09	4.54E-03	50.22 2.30%↑	1.66E-03
	40	44.66	2.80E-03	45.59 2.09%↑	4.20E-03
	50	34.91	3.60E-03	34.43 1.36%↓	8.42E-03
	60	28.46	4.47E-03	28.91 1.57%↑	2.64E-03
Winogrande	10	65.87	5.51E-03	67.42 2.35%↑	3.14E-03
	20	64.31	9.97E-03	66.42 3.27%↑	4.20E-03
	30	63.21	5.59E-03	65.42 3.49%↑	5.50E-03
	40	60.38	9.65E-03	61.47 1.81%↑	5.67E-03
	50	56.17	7.89E-03	56.66 0.86%↑	5.77E-03
	60	51.23	8.90E-03	51.33 0.18%↑	1.06E-02
ARC-c	10	34.51	1.66E-02	34.41 0.29%↓	8.51E-03
	20	33.83	6.95E-03	35.68 5.46%↑	6.61E-03
	30	33.02	7.73E-03	34.75 5.24%↑	1.19E-02
	40	31.25	1.50E-02	33.47 7.10%↑	4.42E-03
	50	26.10	1.58E-02	26.86 2.92%↑	1.24E-02
	60	22.17	2.37E-02	23.64 6.65%↑	6.78E-03
Avg.	10	51.31	7.90E-03	51.97 1.29%↑	4.54E-03
	20	49.86	6.81E-03	51.49 3.27%↑	4.25E-03
	30	48.44	5.95E-03	50.13 3.49%↑	6.35E-03
	40	45.43	9.17E-03	46.85 3.12%↑	4.76E-03
	50	39.06	9.11E-03	39.32 0.66%↑	8.85E-03
	60	33.95	1.23E-02	34.63 1.98%↑	6.66E-03

Table 12: Reasoning benchmarks on recent GQA-based architectures at 10/30/50% head pruning. A single fixed α is used per model without per-task tuning (α=0.9 for Qwen2.5-7B, α=0.7 for Qwen3-8B). HIES outperforms HIS across all pruning ratios on both models.

Pruning	**Method**	**HellaSwag**	**ARC-C**	**ARC-E**	**Winogrande**	**Avg.**
Qwen2.5-7B (GQA, 28Q/4KV, α=0.9)						
10%	HIS	57.25	38.98	59.56	66.46	55.56
	HIES (ours)	57.25	39.66	61.20	66.77	**56.22**
30%	HIS	53.10	35.59	55.74	61.80	51.56
	HIES (ours)	52.85	39.66	57.56	60.38	**52.61**
50%	HIS	45.00	33.56	54.64	53.20	46.60
	HIES (ours)	45.25	38.98	56.65	54.46	**48.84**
Qwen3-8B (GQA, 32Q/8KV, α=0.7)						
10%	HIS	53.65	41.36	57.92	62.12	53.76
	HIES (ours)	54.45	40.68	56.83	63.30	**53.81**
30%	HIS	48.75	35.93	50.64	60.85	49.04
	HIES (ours)	51.70	37.63	53.73	62.83	**51.47**
50%	HIS	41.35	32.88	50.46	57.30	45.50
	HIES (ours)	42.00	35.25	53.55	56.67	**46.87**

CoLA

HIS

	H0	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11
L0	0.04	0.03	0.01	0.01	0.02	0.05	0.06	0.02	0.04	0.04	0.02	0.02
L1	0.05	0.06	0.05	0.05	0.07	0.04	0.07	0.05	0.03	0.04	0.03	0.08
L2	0.13	0.08	0.11	0.07	0.06	0.12	0.10	0.06	0.10	0.11	0.12	0.12
L3	0.09	0.15	0.11	0.11	0.12	0.17	0.10	0.15	0.27	0.19	0.12	0.16
L4	0.20	0.33	0.17	0.09	0.13	0.22	0.13	0.14	0.11	0.14	0.21	0.22
L5	0.24	0.10	0.15	0.21	0.17	0.10	0.13	0.23	0.16	0.37	0.39	0.12
L6	0.21	0.23	0.33	0.34	0.22	0.32	0.29	0.22	0.30	0.23	0.29	0.43
L7	0.26	0.30	0.49	0.15	0.39	0.27	0.15	0.22	0.42	0.43	0.59	0.25
L8	0.22	0.25	0.30	0.23	0.34	0.50	0.29	0.37	0.62	0.43	0.39	0.33
L9	0.15	0.29	0.40	0.26	0.24	0.60	0.24	0.13	0.55	0.11	0.47	0.32
L10	0.89	0.64	0.49	0.75	0.51	0.69	0.35	0.90	0.44	0.62	0.43	0.60
L11	1.00	0.62	0.81	0.77	0.63	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	0.04	0.04	0.02	0.02	0.03	0.09	0.08	0.03	0.06	0.05	0.03	0.02
L1	0.13	0.18	0.12	0.17	0.27	0.13	0.13	0.16	0.11	0.10	0.11	0.20
L2	0.53	0.27	0.20	0.16	0.16	0.25	0.38	0.18	0.31	0.42	0.21	0.50
L3	0.34	0.42	0.23	0.40	0.41	0.43	0.25	0.31	0.40	0.50	0.35	0.40
L4	0.47	0.32	0.23	0.44	0.28	0.52	0.21	0.24	0.31	0.42	0.35	0.27
L5	0.65	0.30	0.28	0.36	0.48	0.31	0.28	0.49	0.78	0.65	0.53	0.25
L6	0.32	0.27	0.39	0.31	0.74	0.60	0.57	0.23	0.74	0.47	0.33	0.72
L7	0.49	0.29	0.77	0.16	0.85	0.42	0.18	0.13	0.52	0.55	0.67	0.44
L8	0.29	0.39	0.75	0.66	0.34	0.40	0.53	0.34	0.23	1.00	0.34	0.56
L9	0.70	0.23	0.22	0.43	0.21	0.32	0.98	0.28	0.54	0.65	0.23	0.57
L10	0.44	0.37	0.33	0.41	0.21	0.19	0.24	0.56	0.26	0.37	0.51	0.32
L11	0.32	0.15	0.39	0.10	0.27	0.46	0.10	0.11	0.89	0.43	0.11	0.29

HIES (Ours)

	H0	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11
L0	0.04	0.05	0.17	0.18	0.02	0.08	0.04	0.02	0.08	0.03	0.21	0.11
L1	0.14	0.27	0.10	0.16	0.35	0.12	0.29	0.15	0.09	0.07	0.15	0.19
L2	0.64	0.31	0.21	0.12	0.16	0.22	0.36	0.16	0.28	0.56	0.17	0.43
L3	0.40	0.34	0.23	0.39	0.34	0.50	0.23	0.25	0.34	0.53	0.32	0.40
L4	0.45	0.28	0.26	0.47	0.24	0.50	0.25	0.31	0.33	0.35	0.35	0.32
L5	0.60	0.36	0.27	0.37	0.40	0.35	0.35	0.50	0.68	0.68	0.53	0.31
L6	0.39	0.33	0.41	0.38	0.67	0.57	0.56	0.28	0.62	0.48	0.39	0.73
L7	0.49	0.34	0.70	0.28	0.83	0.45	0.28	0.26	0.51	0.55	0.63	0.45
L8	0.35	0.40	0.72	0.58	0.39	0.42	0.51	0.39	0.30	0.84	0.40	0.54
L9	0.65	0.32	0.32	0.42	0.32	0.39	0.88	0.38	0.51	0.61	0.32	0.53
L10	0.40	0.42	0.37	0.42	0.31	0.27	0.32	0.52	0.31	0.41	0.46	0.38
L11	0.37	0.25	0.38	0.23	0.33	0.46	0.23	0.23	0.78	0.40	0.23	0.34

QNLI

HIS

	H0	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11
L0	0.02	0.02	0.01	0.01	0.02	0.04	0.06	0.02	0.04	0.03	0.01	0.02
L1	0.23	0.24	0.17	0.26	0.35	0.19	0.16	0.25	0.16	0.16	0.24	0.32
L2	0.81	0.51	0.33	0.21	0.27	0.41	0.67	0.27	0.48	0.65	0.35	0.69
L3	0.58	0.62	0.34	0.63	0.27	0.72	0.38	0.46	0.70	0.84	0.68	0.66
L4	0.55	0.41	0.36	0.52	0.34	0.76	0.34	0.36	0.52	0.48	0.43	0.33
L5	0.71	0.26	0.36	0.36	0.46	0.46	0.36	0.49	0.70	0.79	0.68	0.30
L6	0.41	0.34	0.58	0.27	0.74	0.70	0.54	0.39	0.86	0.60	0.38	0.70
L7	0.66	0.38	0.47	0.20	0.70	0.45	0.16	0.16	0.57	0.62	0.77	0.21
L8	0.28	0.62	0.33	0.51	0.29	0.48	0.51	0.50	0.28	0.65	0.78	0.72
L9	0.44	0.37	0.31	0.45	0.20	0.28	0.56	0.38	0.69	0.19	0.31	0.58
L10	0.31	0.16	0.58	0.43	0.26	0.09	0.24	0.32	0.35	0.49	0.59	0.73
L11	0.31	0.35	0.50	0.30	0.44	0.61	0.29	0.30	1.00	0.44	0.29	0.46

HIES (Ours)

	H0	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11
L0	0.53	0.55	0.15	0.14	0.63	0.56	0.65	0.61	0.51	0.64	0.09	0.32
L1	0.55	0.42	0.64	0.59	0.46	0.64	0.17	0.68	0.69	0.73	0.56	0.76
L2	0.51	0.65	0.66	0.74	0.63	0.78	0.89	0.71	0.79	0.41	0.79	0.90
L3	0.66	0.93	0.68	0.79	0.63	0.63	0.74	0.89	0.96	0.78	0.87	0.78
L4	0.76	0.75	0.67	0.63	0.77	0.79	0.59	0.49	0.72	0.81	0.66	0.50
L5	0.75	0.37	0.66	0.54	0.78	0.58	0.42	0.53	0.77	0.65	0.72	0.45
L6	0.52	0.58	0.70	0.38	0.77	0.74	0.63	0.70	1.00	0.66	0.50	0.63
L7	0.75	0.54	0.56	0.35	0.63	0.55	0.29	0.29	0.65	0.66	0.73	0.38
L8	0.48	0.85	0.48	0.69	0.47	0.61	0.65	0.69	0.55	0.76	0.83	0.77
L9	0.63	0.54	0.56	0.64	0.42	0.49	0.69	0.52	0.80	0.45	0.55	0.84
L10	0.55	0.34	0.76	0.64	0.49	0.29	0.46	0.56	0.51	0.63	0.69	0.88
L11	0.58	0.63	0.73	0.54	0.70	0.77	0.57	0.56	0.98	0.66	0.58	0.67

QQP

HIS

	H0	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11
L0	0.03	0.04	0.01	0.01	0.02	0.04	0.05	0.02	0.04	0.04	0.02	0.02
L1	0.27	0.39	0.27	0.24	0.41	0.19	0.12	0.34	0.22	0.19	0.25	0.38
L2	0.83	0.52	0.20	0.22	0.25	0.38	0.63	0.25	0.38	0.70	0.29	0.57
L3	0.86	0.57	0.27	0.51	0.27	0.63	0.27	0.33	0.69	0.94	0.62	0.59
L4	0.57	0.38	0.14	0.74	0.23	0.75	0.38	0.42	0.29	0.39	0.26	0.30
L5	0.69	0.22	0.25	0.37	0.54	0.43	0.30	0.48	0.64	0.61	0.60	0.25
L6	0.28	0.14	0.35	0.23	0.47	0.37	0.43	0.36	0.40	0.36	0.36	0.51
L7	0.59	0.31	0.45	0.27	0.81	0.29	0.26	0.20	0.21	0.60	0.56	0.27
L8	0.34	0.27	0.20	0.28	0.20	0.32	0.41	0.39	0.22	0.37	0.55	0.61
L9	0.48	0.42	0.33	0.43	0.18	0.40	0.51	0.63	0.76	0.25	0.28	0.42
L10	0.11	0.15	0.40	0.28	0.14	0.09	0.17	0.48	0.27	0.40	0.73	0.41
L11	0.32	0.22	0.38	0.25	0.37	0.77	0.27	0.22	1.00	0.63	0.26	0.59

HIES (Ours)

	H0	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11
L0	0.02	0.03	0.04	0.04	0.01	0.03	0.04	0.01	0.04	0.03	0.05	0.03
L1	0.27	0.39	0.25	0.24	0.41	0.17	0.15	0.32	0.20	0.17	0.24	0.36
L2	0.84	0.51	0.20	0.20	0.24	0.36	0.61	0.24	0.36	0.71	0.28	0.54
L3	0.85	0.54	0.26	0.50	0.25	0.63	0.26	0.31	0.66	0.92	0.60	0.57
L4	0.56	0.36	0.15	0.73	0.22	0.73	0.37	0.42	0.29	0.37	0.26	0.30
L5	0.68	0.23	0.24	0.37	0.52	0.43	0.32	0.48	0.63	0.61	0.59	0.25
L6	0.28	0.15	0.35	0.24	0.46	0.37	0.43	0.35	0.39	0.37	0.37	0.52
L7	0.58	0.31	0.45	0.28	0.80	0.29	0.27	0.21	0.22	0.59	0.55	0.27
L8	0.34	0.27	0.21	0.28	0.20	0.32	0.40	0.38	0.23	0.38	0.54	0.60
L9	0.48	0.41	0.33	0.43	0.19	0.39	0.51	0.62	0.75	0.25	0.28	0.41
L10	0.13	0.17	0.40	0.29	0.16	0.11	0.19	0.47	0.28	0.41	0.72	0.41
L11	0.32	0.22	0.37	0.26	0.36	0.75	0.27	0.23	0.97	0.61	0.26	0.58

Figure 7: Heatmaps of head-importance scores across four GLUE tasks (CoLA, MRPC, QNLI, QQP). Left: HIS; Right: HIES (ours). Rows = layers (L0–L11); columns = heads (H0–H11).

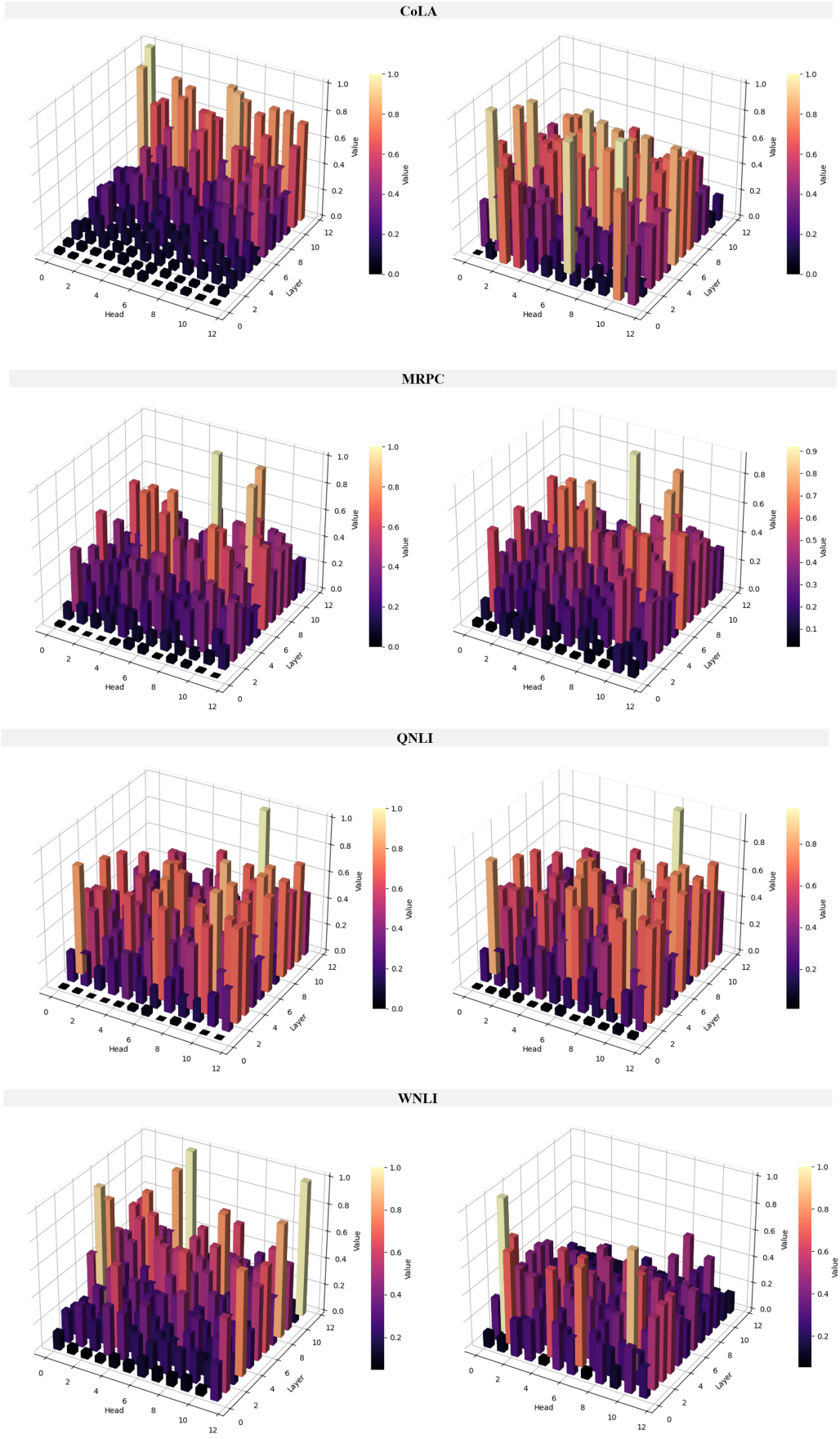


Figure 8: 3D Analysis of Attention Head Importance Scores

CoLA

Pruning Ratio: 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%

	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

	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%

	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

	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%

	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

	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 9: 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.

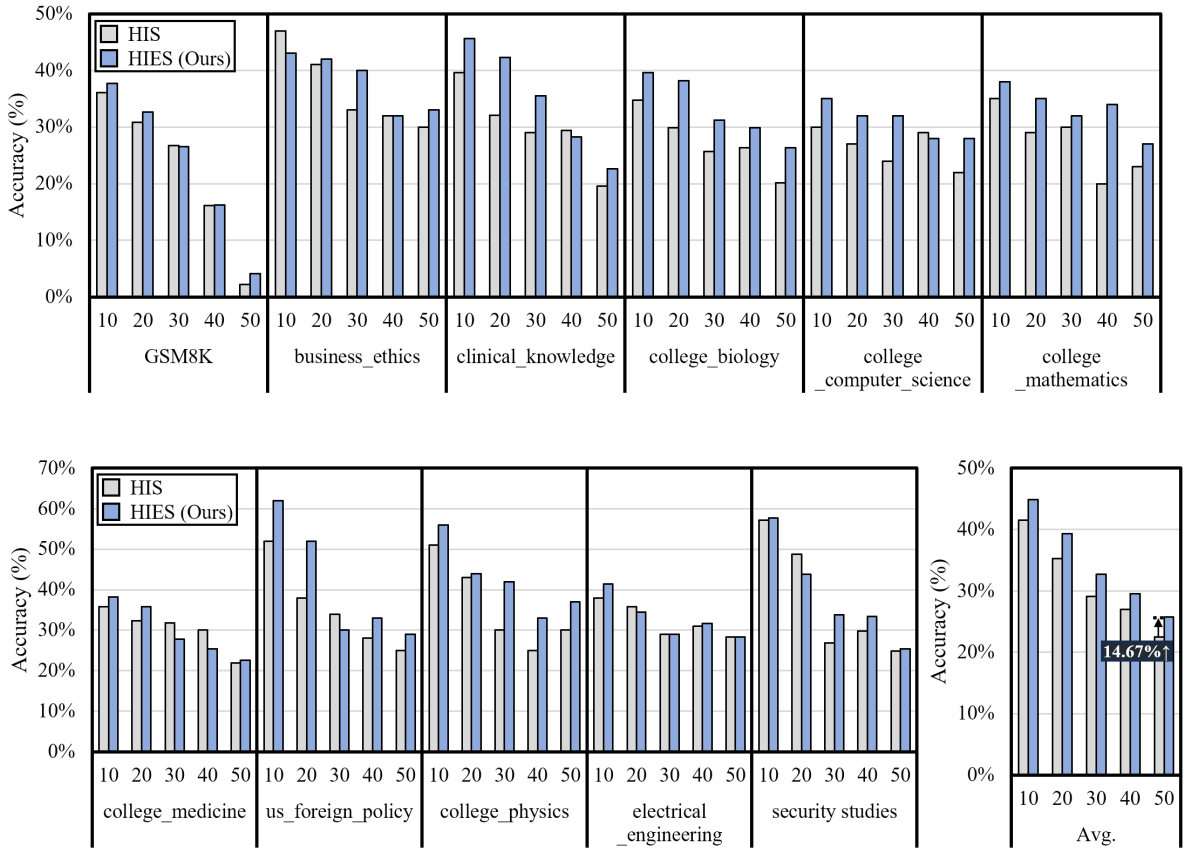


Figure 10: Results on GSM8K (math word-problem reasoning) and MMLU (10-task knowledge reasoning) with LLaMA-2_{7B}.

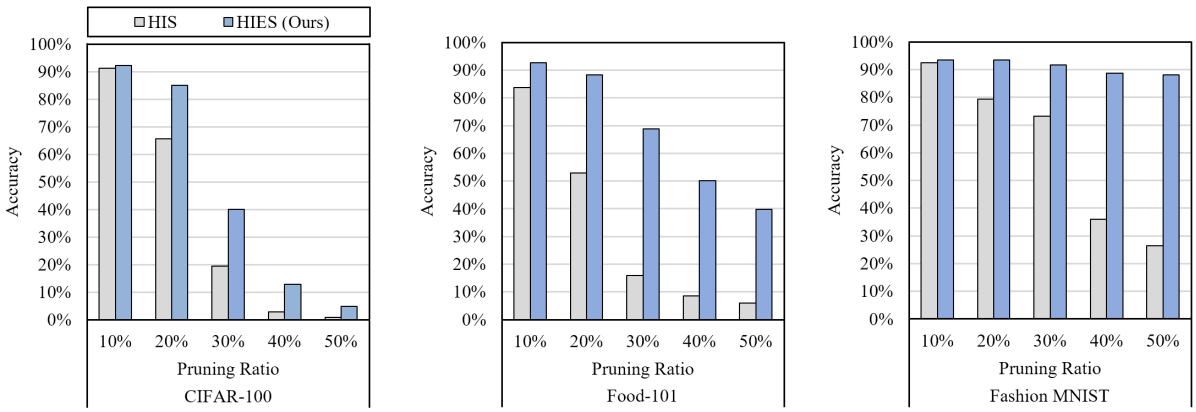


Figure 11: Evaluation of HIES on the image classification benchmarks. HIES consistently outperforms baseline, demonstrating robust and stable performance across downstream tasks.

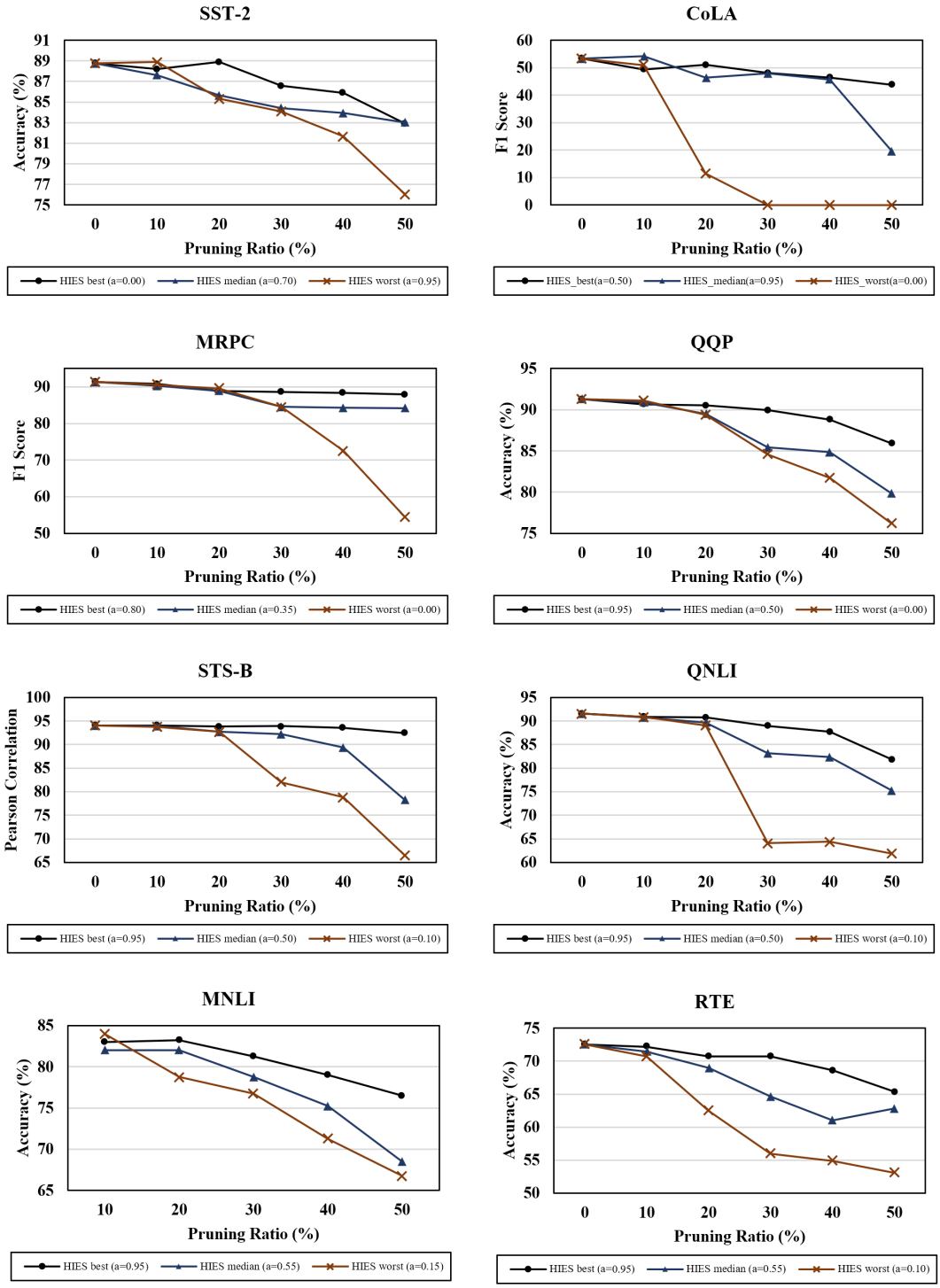


Figure 12: HIES sensitivity to the mixing coefficient α on GLUE. For each task, we sweep α and report three choices—α_{best}, α_{median}, α_{worst}—selected by weighted AUC (wAUC) across pruning ratios. Curves plot performance versus pruning ratio for these three settings.

Table 13: Reasoning benchmarks on additional recent architectures at 10/30/50% head pruning. A single fixed α is used per model without per-task tuning. HIES outperforms HIS across all pruning ratios on all four models.

Pruning	Method	HellaSwag	ARC-C	ARC-E	Winogrande	Avg.
Llama-3.1-8B (GQA, 32Q/8KV, α=0.1)						
10%	HIS	55.30	35.59	67.94	69.85	57.17
	HIES (ours)	55.50	36.27	67.76	69.77	**57.33**
30%	HIS	48.60	32.88	64.12	62.59	52.05
	HIES (ours)	51.50	33.22	66.48	64.33	**53.88**
50%	HIS	31.30	25.08	49.18	53.20	39.69
	HIES (ours)	36.30	28.14	57.74	55.88	**44.51**
DeepSeek-R1-Distill-Llama-8B (GQA, 32Q/8KV, α=0.1)						
10%	HIS	35.00	23.05	37.16	51.07	36.57
	HIES (ours)	35.50	25.42	41.53	51.78	**38.56**
30%	HIS	28.40	21.02	32.97	49.49	32.97
	HIES (ours)	30.30	22.03	34.24	51.22	**34.45**
50%	HIS	25.30	20.68	27.14	48.07	30.30
	HIES (ours)	25.60	24.75	32.24	48.93	**32.88**
OLMo-3-7B (MHA, 32Q/32KV, α=0.6)						
10%	HIS	50.40	35.59	61.02	61.88	52.22
	HIES (ours)	50.80	33.90	63.21	62.51	**52.60**
30%	HIS	46.60	33.22	59.93	59.51	49.81
	HIES (ours)	47.50	33.90	61.38	61.48	**51.07**
50%	HIS	34.50	27.46	55.92	56.27	43.54
	HIES (ours)	38.50	29.15	55.01	55.25	**44.48**
Qwen3-1.7B (GQA, 16Q/8KV, α=0.9)						
10%	HIS	43.10	31.86	59.02	57.14	47.78
	HIES (ours)	43.20	32.20	59.93	56.35	**47.92**
30%	HIS	35.90	27.12	56.10	53.20	43.08
	HIES (ours)	38.20	31.86	61.02	55.56	**46.66**
50%	HIS	25.90	14.58	28.96	50.67	30.03
	HIES (ours)	25.80	23.05	34.97	49.64	**33.37**
Qwen2.5-3B (GQA, 16Q/2KV, α=0.4)						
10%	HIS	50.40	33.90	66.48	64.56	53.84
	HIES (ours)	51.30	37.29	67.58	64.01	**55.04**
30%	HIS	29.30	33.22	47.18	51.14	40.21
	HIES (ours)	46.50	30.17	60.29	62.51	**49.87**
50%	HIS	25.10	20.34	28.60	50.43	31.12
	HIES (ours)	28.20	24.07	37.34	50.28	**34.97**

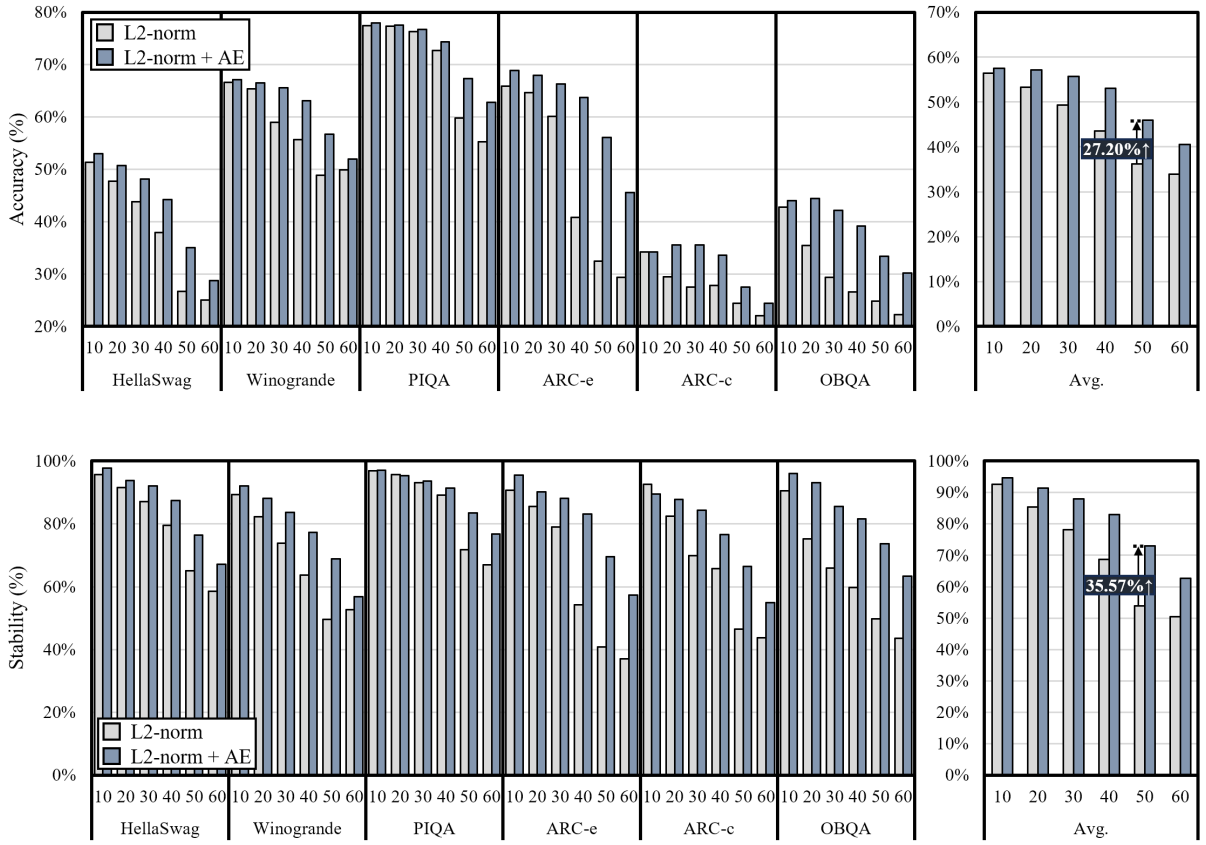


Figure 13: Accuracy and stability improvements when combining AE with gradient L2-norm–based importance scores, compared to using the L2-norm alone. We conduct experiments on LLaMA-2_{7B} for HellaSwag, Winogrande, PIQA, ARC-e, ARC-c, and OBQA.