

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

Minsik Choi^{1,*} Hyegang Son^{1,*} Changhoon Kim^{2,†} Young Geun Kim^{1,†}

¹Korea University ²Soongsil University

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 a 1.13% improvement in model quality and 3.52% improvement in stability over the HIS-only method on average, substantially reducing the accuracy and stability cost of aggressive compression. 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., 2024b). 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 identify 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., 2024; 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

*Equal contribution.

†Corresponding authors:

kch@ssu.ac.kr, younggeun_kim@korea.ac.kr

that different layers require distinct attention behaviors (Ben Artzy and Schwartz, 2024). Lacking such layer-specific characteristics often results in imbalanced pruning—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 pruning yields a 1.13% improvement in model quality and 3.52% improvement in stability over the HIS-only method on average. 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 memory requirement¹ while

¹Note that the memory benefit of head pruning is architecture-dependent. Under multi-head attention (MHA), pruning a head removes its key and value projections and corresponding KV-cache footprint. On the other hand, under grouped-query attention (GQA), pruning individual query heads does not reduce the shared KV cache unless an entire group is removed.

preserving the layer topology, thereby simplifying checkpoint compatibility and serving integration (Li et al., 2021; Behnke and Heafeld, 2020). Consequently, large-scale studies adopt head-level pruning as a practical axis in LLM compression pipelines (Jaradat et al., 2024; Muralidharan et al., 2024)². 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:

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

where $\mathcal{D}$ denotes an input sample drawn from the data distribution $\mathcal{D}$, $\mathcal{L}(x)$ is the loss for sample x , and $A_h(x)$ 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 $\text{AE}_h = (H(p^{(h)})) = -\sum_{i=1}^n p_i^{(h)} \log p_i^{(h)}$, 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 *prediction agreement* (equivalently, fidelity to the original model; its complement is prediction churn) between the pruned and original models: $\text{Stability} = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \mathbf{1}[y_{\text{pruned}}(x) = y_{\text{original}}(x)]$.

3 Motivation

Pruning Transformer models is most commonly driven by gradient-based criteria, such as HIS and

²For more detailed discussions on related work, please refer to Appendix A.

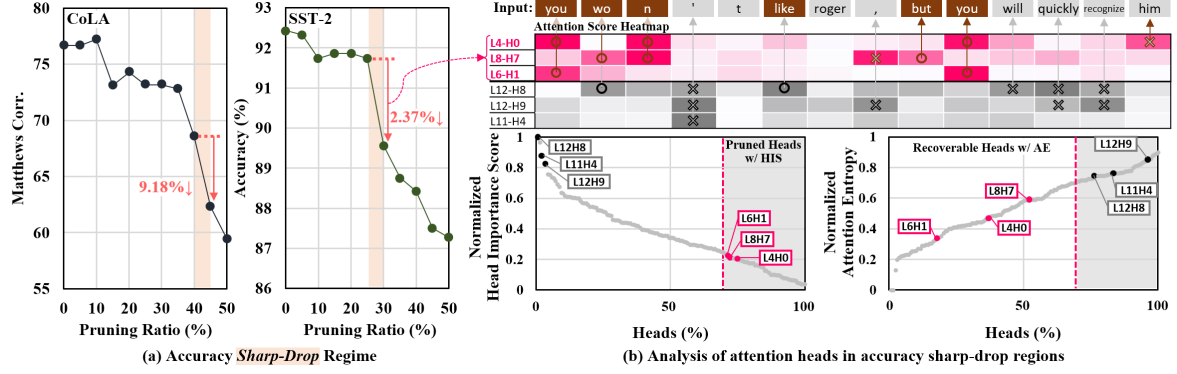


Figure 1: Analysis of accuracy degradation and head behaviors under HIS-based pruning. In our diagnostic study, we analyze the phenomena of pruning by HIS on BERT, focusing on detailed attention head behaviors during inference. (a) Accuracy curves of HIS-based pruning on CoLA and SST-2. The bold color 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.

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.

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., 2026a). 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 attention 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:

$$\text{HIES}_h = \alpha \widehat{\text{HIS}}_h + (1 - \alpha)(1 - \widehat{\text{AE}}_h), \quad (2)$$

where $\alpha \in [0, 1]$ is a mixing coefficient³.

This form is motivated by 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 $1 - \text{AE}$. A weighted-sum scalarization of the two recovers the full set of Pareto-optimal trade-offs under convex conditions (Boyd and Vandenberghe, 2004). The loss-increase bound (Lemma 4.1) identifies $\widehat{\text{HIS}}_h$ as the quantity a pruning criterion should preserve, while the stability-based generalization-gap bound (Eq. (6); Bousquet and Elisseeff, 2002) shows that the retained heads’ aggregate entropy deficit controls the generalization gap; the complementary term $1 - \widehat{\text{AE}}_h$ is motivated by the diagnostic analysis in Section 3. The two terms’ gradient directions are orthogonal in expectation (Lemma 4.3), so the two signals combine linearly without gradient conflict (Yu et al., 2020), and Remark 4.2 notes that, because this surrogate is modular, top- k selection by HIES exactly minimizes it under the cardinality constraint.

³The α -selection protocol is described in Section 5.1; sensitivity is analyzed in Appendix D.8, and fixed- α results across model families are presented in Appendix D.10.

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 $[0, 1]$: $\widehat{\text{HIS}}_h = \frac{\text{HIS}_h - \min(\text{HIS})}{\max(\text{HIS}) - \min(\text{HIS})}$, $\widehat{\text{AE}}_h = \frac{\text{AE}_h - \min(\text{AE})}{\max(\text{AE}) - \min(\text{AE})}$. This distribution-agnostic normalization improves cross-criterion interpretability; lower normalized scores denote higher pruning priority. We prefer min–max scaling to z -score standardization because its bounded co-domain $[0, 1]$ makes α an interpretable mixing weight between two commensurate quantities, whereas z -scores are unbounded and signed, so a fixed α would not correspond to a fixed trade-off. Broad empirical comparisons of scaling techniques report no universally optimal choice (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. Retaining heads with high HIES targets both terms of this decomposition through a single linear surrogate. Conceptually, this analysis formalizes importance-based selection into a principled framework and clarifies why HIES is a sensible criterion.

4.2.1 Loss-Increase Control via Head Importance (HIS)

Setup. Let n be the sequence length, $|\mathbf{H}|$ the number of heads, d_v the value dimension per head, and $d = |\mathbf{H}| d_v$ 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 $A_h = \text{softmax}\left(\frac{Q_h K_h^\top}{\sqrt{d_k}}\right) V_h$, where $Q_h = X W_h^Q \in \mathbb{R}^{n \times d_k}$, $K_h = X W_h^K \in \mathbb{R}^{n \times d_k}$, and $V_h = X W_h^V \in \mathbb{R}^{n \times d_v}$ 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 $\mathbf{y} = \text{Concat}(A_1, \dots, A_{|\mathbf{H}|}) \in \mathbb{R}^{n \times d}$ as the pre-projection representation, which

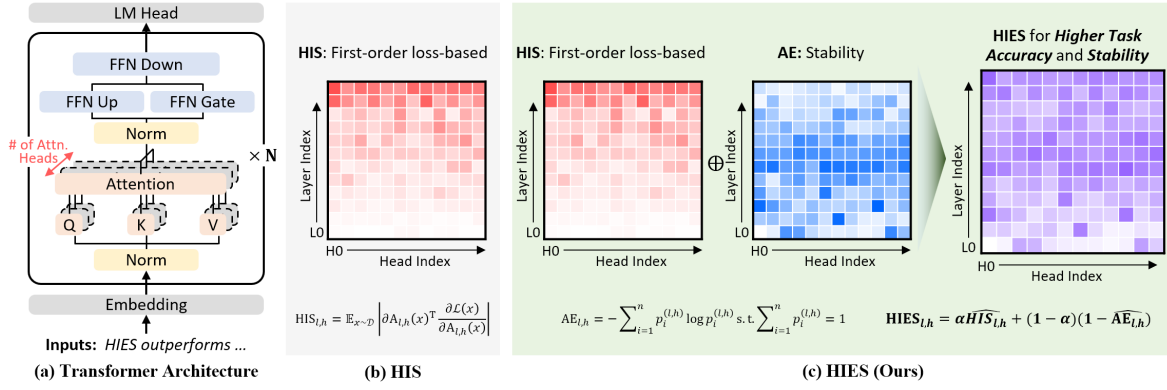


Figure 2: Design overview of Head Importance-Entropy Score (HIES). Darker cells correspond to values closer to 1, while lighter cells correspond to values closer to 0. Note the heatmap utilizes each metric across attention heads in BERT on the CoLA dataset.

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 $\delta \mathbf{y} = \text{Concat}(\delta A_1, \dots, \delta A_{|H|})$. Formal definitions and implementation notes are deferred to Appendix B.1.2.

Head Importance Score (HIS). We define

$$\begin{aligned} \text{HIS}_h &:= \mathbb{E}_{x \sim \mathcal{D}} \left| \frac{\partial \mathcal{L}(x)}{\partial m_h} \right| = \mathbb{E}_{x \sim \mathcal{D}} \left| \langle \nabla_{A_h(x)} \mathcal{L}(x), A_h(x) \rangle_F \right| \\ &= \mathbb{E}_{\mathcal{D}} \left[\left| \langle \nabla_{A_h} \mathcal{L}, A_h \rangle_F \right| \right]. \end{aligned} \quad (3)$$

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

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

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

Equivalently, writing the bound with the curvature made explicit,

$$\Delta \mathcal{L} \leq \sum_{h=1}^{|H|} (1 - m_h) \text{HIS}_h + \frac{\beta_y}{2} \sum_{h=1}^{|H|} (1 - m_h) \|A_h\|_F^2. \quad (4)$$

Remark. β_y is the curvature of the loss with respect to the pre-projection representation $\mathbf{y}$. Since the map from $\mathbf{y}$ to the classifier logits is not a single linear layer, we do not bound β_y in closed form; the standard cross-entropy curvature bounds ($\frac{1}{4}$ binary, $\frac{1}{2}$ multiclass) apply at the logits and are stated in Appendix B.1.4 for reference.

Implication for pruning. Eq. (4) shows that, for a fixed pruning fraction $\rho = \frac{1}{|H|} \sum_h (1 - m_h)$, selecting heads with the smallest HIS_h minimizes the dominant first-order term, while the quadratic term is controlled by β_y 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 $A_h = \text{Attn}_h V_h$, HIS_h in Eq. (3) 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 $\alpha_t^{(h)}(x) \in \Delta^{n(x)-1}$ 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)

$$\begin{aligned} \text{AD}_h(x) &:= \frac{1}{n(x) \log n(x)} \times \sum_{t=1}^{n(x)} \left(\log n(x) - H(\alpha_t^{(h)}(x)) \right) \\ &= 1 - \text{AE}_h(x) \in [0, 1]. \end{aligned} \quad (5)$$

Main bound (loss–entropy link). Let $M := \max_h \max_j \|V_h(j, :)\|_2$ and $C_{\text{AE}} := \sqrt{8} M \sqrt{|H|(1 - \rho) \log n}$ for a representative effective length n . For the pruned model $f_{S,m}$, the expected generalization gap ($\mathcal{G}$) satisfies

$$\begin{aligned} \mathcal{G} &:= \mathbb{E}_{\mathcal{D}}[\ell(f_{S,m}(x), y)] - \mathbb{E}_S[\ell(f_{S,m}(x), y)] \\ &\leq 2L_\ell C_{\text{AE}} \sqrt{\sum_{h=1}^{|H|} m_h \mathbb{E}_S[\overline{\text{AD}}_h(x)]} + \frac{B}{N}. \end{aligned} \quad (6)$$

Interpretation. Eq. (6) bounds the generalization gap of the *retained* sub-network: pruned heads are absent from $f_{S,m}$ and contribute nothing, so the perturbation is carried by the heads with $m_h = 1$ and is controlled by their aggregate entropy deficit. The bound therefore constrains how much generalization the surviving heads can lose, but it does not by itself prescribe which heads to remove; our use of $1 - \text{AE}_h$ as a selection signal is motivated empirically in Section 3 and Appendix D.8. Proof details, operator-norm assumptions, and variable-length handling are deferred to Appendix B.2.

4.2.3 Selection Surrogate and HIES Minimization

Composite Selection Surrogate. Given the HIES above, we define the linear *selection surrogate*

$$\mathcal{J}(m) := \sum_{h=1}^{|H|} (1 - m_h) \text{HIES}_h \quad (7)$$

Eq. (7) is a *heuristic linear scalarization* of the two analyses above rather than an upper bound on the risk: the generalization term of Eq. (6) enters under a square root and carries explicit constants, and the min–max normalization makes HIES_h invariant to a rescaling of $\mathcal{L}$ to which $\Delta\mathcal{L}$ is not. We adopt it because it is the simplest score that is monotone in both signals and admits an exact top- k solution (Remark 4.2).

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

$$\min_{m \in \{0,1\}^{|H|}} \sum_{h=1}^{|H|} (1 - m_h) \text{HIES}_h \quad \text{s.t.} \quad \sum_{h=1}^{|H|} m_h = k. \quad (8)$$

Remark 4.2 (Exact minimizer of the selection surrogate). Because (7) is modular in m , selecting the k heads with the largest HIES values (equivalently, pruning the $|H| - k$ heads with the smallest HIES values) exactly minimizes (7) subject to (8).

4.2.4 Orthogonality and Complementarity

Lemma 4.3 (Orthogonality). *Let $u_h := \text{sign}(\alpha^{(h)\top} g_h) g_h$, $v_h := \mathbf{1} + \log \alpha^{(h)}$, $\tilde{u}_h := P u_h$, $\tilde{v}_h := P v_h$, where $P := I - \frac{1}{n} \mathbf{1}\mathbf{1}^\top$ projects onto $\{w : \mathbf{1}^\top w = 0\}$. Assume $\text{Cov}(\tilde{u}_h, \tilde{v}_h) = 0$ (the cross-covariance matrix is zero). If, in addition, either $\mathbb{E}_S[\tilde{u}_h] = 0$ or more generally $\langle \mathbb{E}_S[\tilde{u}_h], \mathbb{E}_S[\tilde{v}_h] \rangle = 0^4$, then*

$$\mathbb{E}_S[\langle \tilde{\nabla}_{\alpha^{(h)}} \text{HIS}_h, \tilde{\nabla}_{\alpha^{(h)}} \text{AE}_h \rangle] = 0, \quad (9)$$

i.e., the two gradient directions are orthogonal in expectation⁵.

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 last). At the channel and block level, **LLM-Pruner** (Ma et al., 2023) provides two variants restricted to attention 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

⁴We provide an empirical sanity check supporting the assumption, which is consistently validated on both BERT_{base} and LLaMA-2_{7B} in Appendix D.1.

⁵Detailed derivations and preliminaries are deferred to Appendix B.4.

Table 1: Experimental results with BERT_{base} on natural language understanding task. We report percentage improvements over the Random baseline in **blue**. For the two correlation-based tasks (CoLA, Matthews; STS-B, Pearson) we report the correlation rescaled to $[0, 100]$ as $(c + 1)/2 \times 100$, so that every column shares a common range; the raw correlation is recovered as $2v/100 - 1$. Appendix Table 9 uses the same rescaled axis for CoLA but reports STS-B as a raw Pearson correlation.

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%	BERT _{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
	AD	92.55	75.48	90.56	91.10	93.65	90.87	83.50	70.40	86.01
	HIS	91.74	77.21	<u>90.65</u>	<u>91.23</u>	94.04	90.63	84.04	71.84	86.42
	L2	90.37	74.00	83.78	69.23	83.12	60.81	67.75	49.82	72.36
	LLM-Pruner (Channel)	91.97	76.05	79.19	83.53	93.06	87.39	80.46	67.51	82.40
	LLM-Pruner (Block)	91.06	76.69	89.83	84.96	93.77	87.42	82.03	64.62	83.80
	SliceGPT (w/o tune)	51.38	49.86	81.46	63.18	58.92	53.91	36.84	49.46	55.63
	SliceGPT (w/ tune)	86.47	61.87	82.30	88.28	62.47	83.45	77.34	54.51	74.59
	HIES (ours)	92.66	75.48	91.04	91.93	<u>94.00</u>	91.03	84.04	71.84	86.50
30%	Random	90.29	69.02	85.27	84.00	92.90	79.80	75.15	60.29	79.59
	AD	86.58	50.00	84.53	84.57	81.94	67.82	77.00	56.68	73.64
	HIS	89.56	73.17	89.37	89.95	93.82	89.04	82.25	68.59	84.47
	L2	86.58	67.52	81.58	64.83	76.52	51.09	56.00	50.90	66.88
	LLM-Pruner (Channel)	88.53	70.36	86.89	81.23	92.50	67.38	66.67	64.26	77.23
	LLM-Pruner (Block)	88.99	<u>73.93</u>	84.76	80.09	<u>93.40</u>	82.68	78.33	66.79	81.12
	SliceGPT (w/o tune)	50.80	53.29	78.05	63.18	54.64	53.18	34.90	51.99	55.00
	SliceGPT (w/ tune)	83.49	60.14	81.80	85.80	60.89	67.36	75.50	54.87	71.23
	HIES (ours)	91.86	74.97	<u>88.81</u>	90.37	93.89	89.13	82.50	70.04	85.20
50%	Random	78.74	61.02	72.53	66.25	91.40	67.53	67.32	53.79	69.82
	AD	82.91	50.00	54.50	76.18	75.00	68.94	68.00	55.96	66.44
	HIS	87.27	59.48	<u>86.52</u>	85.91	92.61	82.68	<u>78.67</u>	<u>62.82</u>	79.50
	L2	82.80	60.98	85.30	64.83	69.76	50.54	44.42	47.29	63.24
	LLM-Pruner (Channel)	86.47	61.64	83.92	81.47	89.74	60.66	67.42	62.82	74.27
	LLM-Pruner (Block)	87.84	70.09	83.84	78.80	92.69	72.60	73.60	61.73	77.65
	SliceGPT (w/o tune)	50.92	52.79	81.22	63.19	47.70	50.98	34.93	50.90	54.08
	SliceGPT (w/ tune)	83.49	57.45	81.37	82.16	55.05	65.79	71.70	51.62	68.58
	HIES (ours)	90.71	<u>68.52</u>	86.80	<u>85.73</u>	<u>92.65</u>	82.68	79.00	65.34	81.43

Table 2: Stability results with BERT_{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
	L2	95.18	76.41	26.72	68.63	21.87	60.33	72.44	39.71	57.66
	AD	98.51	96.36	93.63	97.80	89.07	97.25	96.15	94.95	95.47
	LLM-Pruner	97.02	85.81	75.25	85.40	39.53	90.39	84.98	86.28	80.58
	HIES (ours)	<u>98.17</u>	<u>96.36</u>	<u>94.12</u>	97.80	<u>96.67</u>	97.25	<u>95.74</u>	93.50	96.20
30%	HIS	94.38	90.03	90.03	94.77	78.93	92.95	91.56	87.36	90.00
	L2	90.94	83.13	26.72	63.90	20.73	50.39	59.23	42.24	54.66
	AD	88.53	81.21	78.68	87.90	24.53	63.04	79.07	69.68	71.58
	LLM-Pruner	92.66	63.09	84.80	82.37	62.13	68.70	68.29	70.04	74.01
	HIES (ours)	97.13	93.38	<u>89.46</u>	94.97	86.67	93.23	91.56	88.81	91.90
50%	HIS	90.02	40.84	83.33	88.80	69.13	85.54	81.47	78.70	77.23
	L2	86.47	82.26	26.72	63.90	20.40	49.81	45.58	34.30	51.18
	AD	52.18	81.21	51.72	72.33	21.87	73.49	68.49	62.82	60.51
	LLM-Pruner	89.68	41.32	78.19	83.27	44.13	61.58	67.52	75.81	67.69
	HIES (ours)	95.07	83.13	84.56	<u>88.27</u>	75.20	85.54	<u>78.29</u>	81.23	83.91

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.

Mixing coefficient. For HIES, we select α once per model on a 32-example HellaSwag calibration set held out from the evaluation subset (Appendix D.3), based on the highest mean calibration accuracy across pruning ratios, and then keep it fixed across tasks and pruning ratios. Because HIS and AE come from one shared calibration pass, the search only recombines the two cached scores and adds no model passes: requiring only 66 ms on LLaMA-2_{7B} and 9 ms on BERT.

5.2 Main Results

We evaluate HIES using two key metrics: model quality and stability (Table 1), with additional evaluation on recent GQA-based decoder models in Section 5.3.4⁶. As reported in Table 1, averaged over the three pruning ratios HIES improves model quality by 8.21% over Random pruning and by 1.13% over HIS. Table 2 further demonstrates a 3.52% average stability gain over HIS. Notably, on CoLA at a pruning ratio of 50%, HIES achieves gains of up to 15.2% in model quality and 2.04× in

⁶Experimental details are provided in Appendix C. We also conduct sensitivity analysis across multiple random seeds and observe consistent improvements for both BERT_{base} and LLaMA-2_{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.

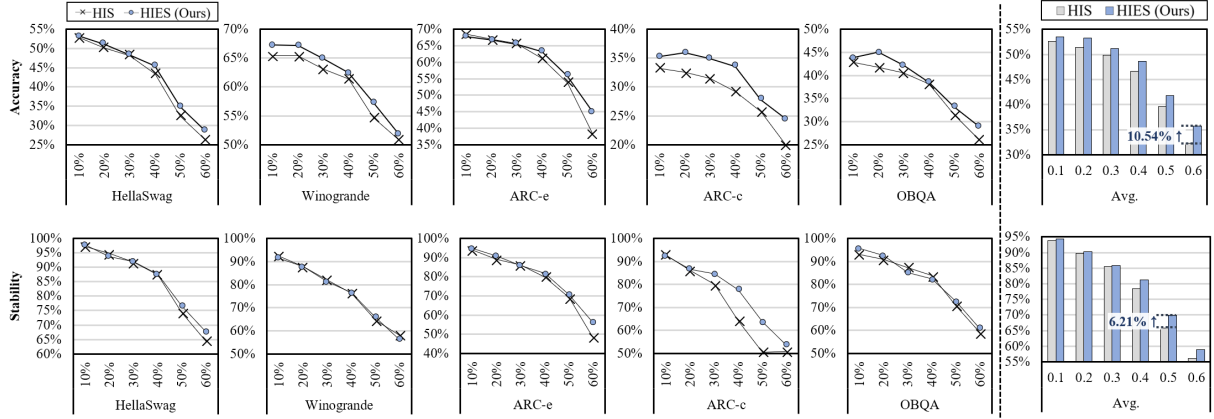


Figure 3: Comparison of HIES (Ours) and HIS on HellaSwag, Winogrande, ARC-e, ARC-c, and OBQA across pruning ratios from 10% to 60% (x-axis). The top row reports task accuracy, while the bottom row reports stability. The rightmost panels summarize the mean over tasks.

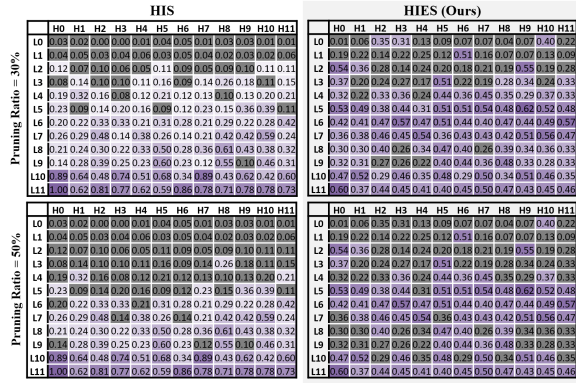


Figure 4: Head-level pruning patterns on the CoLA dataset across pruning ratios from 30% to 50% pruning ratios. Pruned heads are shaded in gray.

stability relative to HIS. These results corroborate our theoretical analysis, demonstrating that HIES preserves end-task performance while markedly enhancing robustness; HIES achieves a $1.42\times$ inference speedup at a 50% pruning ratio and a sequence length of 4096, with potentially larger gains for longer contexts. Detailed efficiency analysis is provided in Appendix E.

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., $\geq 30\%$). At lower ratios (e.g., $\leq 10\%$), 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.

Pruning heatmap analysis provides empirical support for this distinction. As shown in Figure 4, 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.

5.3.2 Scalability to Larger Models

On LLaMA-2_{7B}, we evaluate pruning on HellaSwag, Winogrande, ARC-e, ARC-c, and OBQA, comparing HIES with HIS across pruning ratios of 10–60% (Figure 3). 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

Table 3: Experimental results with ViT_{Large} on image classification benchmarks and LLaVA-1.5_{7B} on multi-modal tasks. Relative improvements of HIES (Ours) over HIS are shown in blue.

Vision Transformer Model (ViT _{Large})							Vision-Language Model (LLaVA-1.5 _{7B})					
Image Classification							Visual Question Answering		Complex Multimodal			
	ImageNet1k		CIFAR-100		HIS	Avg. HIES (Ours)	VizWiz-VQA		MM-Vet		HIS	Avg. 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% -2.76%
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% 9.76%
30%	6.20%	55.40%	19.40%	40.00%	12.80%	47.70% 272.66%	50%	32.67%	39.67%	24.60%	25.20%	28.64% 32.43% 13.23%

Method	Qwen2.5 _{7B}			Qwen3 _{8B}		
	10%	30%	50%	10%	30%	50%
Random	54.11	44.37	33.07	54.59	47.33	39.99
AD (1-AE)	32.77	32.65	31.57	51.69	35.63	32.94
L2	36.81	33.80	32.48	52.09	47.83	38.19
HIS	55.56	51.56	46.60	53.76	49.04	45.50
AMP	54.00	50.32	40.46	54.97	51.24	44.23
HARP	54.41	50.12	43.81	52.87	48.75	41.11
GISP	55.02	50.04	41.59	54.51	52.85	44.11
HIES (ours)	56.22	52.61	48.84	53.81	51.47	46.87

Table 4: Experimental results with Qwen2.5_{7B} and Qwen3_{8B} compared to concurrent works. Note that we report average zero-shot accuracy under fixed- α head pruning.

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 ViT_{Large} and visual-language reasoning with LLaVA-1.5_{7B}.

First, across attention-based variants, HIES consistently shifts the sharp accuracy-drop region to more aggressive sparsity levels: on ViT_{Large} at 20% sparsity, HIES raises average accuracy from 55.20% to 82.70%, a 49.82% relative gain over HIS.

Second, the gains carry over to multi-modal tasks: on VizWiz-VQA and MM-Vet, HIES improves accuracy by 13.23% 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 (Yang et al., 2024a), Qwen3 (Yang et al., 2025), LLaMA-3.1 (Grattafiori et al., 2024), DeepSeek-R1 (Guo et al., 2025), and OLMo-3 (Team Olmo et al., 2025) families, spanning Grouped-Query Atten-

tion (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.10.

5.3.4 Comparison with Concurrent Works

Table 4 compares HIES against recent structured pruning baselines (Mugnaini et al., 2025; Liu and Liu, 2025; Wang et al., 2026b) on HellaSwag, WinoGrande, ARC-e, and ARC-c. In the aggressive pruning ratio (i.e., 50%), HIES achieves the best performance on both models (48.84 and 46.87), outperforming the strongest baseline by +5.0 and +2.6, respectively. While AMP (Mugnaini et al., 2025) and GISP (Wang et al., 2026b) slightly outperform HIES at low sparsity on Qwen3-8B, their performance drops sharply at 50% pruning on Qwen2.5_{7B} (40.46 and 41.59). In contrast, HIES mitigates this sharp accuracy degradation at high sparsity by preserving low-entropy specialized heads through its entropy-aware criterion.

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 outperforms HIS and other baselines in accuracy and inference-time stability under aggressive pruning ($\geq 30\%$), while performing comparably at mild ratios ($\leq 10\%$). 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 at aggressive pruning ratios across diverse Transformer architectures and tasks, its scope is bounded in three

ways. First, head pruning is a composable compression axis rather than a standalone solution: because feed-forward layers dominate Transformer computation, pruning 50% of attention heads removes only about 20% of total FLOPs (Appendix E.1), and the memory benefit is further limited under GQA, where pruning individual query heads leaves the shared KV cache intact. HIES is therefore best deployed alongside orthogonal techniques rather than pushed to extreme head sparsity alone. Second, 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. Third, 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.

Acknowledgments

This work was supported by National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-24534857 and RS-2026-25522655), Institute of Information & Communications Technology Planning & Evaluation(IITP)-ICT Creative Consilience Program grant funded by the Korea government (MSIT) (IITP-2026-RS-2020-II201819, 5%), IITP-ITRC(Information Technology Research Center) grant funded by the Korea government (MSIT) (IITP-2026-RS-2023-00260091, 5%), IITP grant funded by the Korea government (MSIT) (RS-2024-00398353, Development of Countermeasure Technologies for Generative AI Security Threats), and AI Seoul Tech Research Support Program of the Seoul Future Foundation. We also thank the members of the Korea University Intelligent Computer Architecture & Systems research Lab and the anonymous reviewers for their helpful feedback.

References

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sang-

hai. 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, Singapore. Association for Computational Linguistics.

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.

Saleh Ashkboos, Maximilian L. Croci, Marcelo Genari do Nascimento, Torsten Hoeffer, and James Hensman. 2024. [SliceGPT: Compress Large Language Models by Deleting Rows and Columns](#). In *The Twelfth International Conference on Learning Representations*.

Yonatan Ashlag, Uri Koren, Mirco Mutti, Esther Derman, Pierre-Luc Bacon, and Shie Mannor. 2025. [State Entropy Regularization for Robust Reinforcement Learning](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 21018–21046. Curran Associates, Inc.

Haoli Bai, Wei Zhang, Lu Hou, Lifeng Shang, Jing 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, Online. Association for Computational Linguistics.

Anna Bair, Hongxu Yin, Maying Shen, Pavlo Molchanov, and Jose M. Alvarez. 2024. [Adaptive Sharpness-Aware Pruning for Robust Sparse Networks](#). In *The Twelfth International Conference on Learning Representations*.

Maximiliana Behnke and Kenneth Heafield. 2020. [Losing Heads in the Lottery: Pruning Transformer Attention in Neural Machine Translation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2664–2674, Online. Association for Computational Linguistics.

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, Miami, Florida, US. Association for Computational Linguistics.

Homanga Bharadhwaj, Kevin Xie, and Florian Shkurti. 2020. Model-Predictive Control via Cross-Entropy and Gradient-Based Optimization. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pages 277–286. PMLR.

Jose Blanchet, Peng Cui, Jiajin Li, and Jiashuo Liu. 2024. Stability Evaluation through Distributional

- Perturbation Analysis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 4140–4159. PMLR.
- Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. [Food-101 – Mining Discriminative Components with Random Forests](#). In *Computer Vision – ECCV 2014*, volume 8694 of *Lecture Notes in Computer Science*, pages 446–461, Cham. Springer.
- Olivier Bousquet and André Elisseeff. 2002. Stability and Generalization. *Journal of Machine Learning Research*, 2(Mar):499–526.
- Stephen Boyd and Lieven Vandenbergh. 2004. [Convex Optimization](#). Cambridge University Press, Cambridge, UK.
- 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, Vienna, Austria. Association for Computational Linguistics.
- 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](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5360–5376, Suzhou, China. Association for Computational Linguistics.
- Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. [What Does BERT Look at? An Analysis of BERT’s Attention](#). In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.
- 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, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*.
- Lucas B. V. de Amorim, George D. C. Cavalcanti, and Rafael M. O. 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, Minneapolis, Minnesota. Association for Computational Linguistics.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. [An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale](#). In *International Conference on Learning Representations*.
- Angela Fan, Edouard Grave, and Armand Joulin. 2020. [Reducing Transformer Depth on Demand with Structured Dropout](#). In *International Conference on Learning Representations*.
- Tamer Ghattas, Michael Hassid, and Roy Schwartz. 2025. [On Pruning State-Space LLMs](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18800–18814, Suzhou, China. Association for Computational Linguistics.
- Mitchell A. 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, Online. Association for Computational Linguistics.
- 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 incentivizes reasoning in LLMs through reinforcement learning](#). *Nature*, 645(8081):633–638.
- 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 (CVPR)*, 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 J. Dally. 2015. Learning both Weights and Connections for Efficient Neural Network. In *Advances in Neural*

- Information Processing Systems*, volume 28, pages 1135–1143. Curran Associates, Inc.
- Moritz Hardt, Ben Recht, and Yoram Singer. 2016. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1225–1234, New York, New York, USA. PMLR.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring Massive Multitask Language Understanding](#). In *International Conference on Learning Representations*.
- Torsten Hoeftler, 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 8360–8378, Suzhou, China. Association for Computational Linguistics.
- Lu Hou, Zhiqi Huang, Lifeng Shang, Xin Jiang, Xiao Chen, and Qun Liu. 2020. DynaBERT: Dynamic BERT with Adaptive Width and Depth. In *Advances in Neural Information Processing Systems*, volume 33, pages 9782–9793. Curran Associates, Inc.
- 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. 2009. Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto, Toronto, ON, Canada.
- 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](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4163–4181, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- François Lagunas, Ella Charlaix, Victor Sanh, and Alexander M. Rush. 2021. [Block Pruning For Faster Transformers](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10619–10629, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- 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, Online. Association for Computational Linguistics.
- Bingbing Li, Zigeng Wang, Shaoyi Huang, Mikhail Bragin, Ji Li, and Caiwen Ding. 2023a. [Towards Lossless Head Pruning through Automatic Peer Distillation for Language Models](#). In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 5113–5121. International Joint Conferences on Artificial Intelligence Organization.
- Jiaoda Li, Ryan Cotterell, and Mrinmaya Sachan. 2021. [Differentiable Subset Pruning of Transformer Heads](#). *Transactions of the Association for Computational Linguistics*, 9:1442–1459.
- 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 *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 20336–20350. PMLR.
- Felipe Tomazelli Lima and Vinicius M. A. Souza. 2023. [A Large Comparison of Normalization Methods on Time Series](#). *Big Data Research*, 34:100407.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 3214–3252.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. [Visual Instruction Tuning](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.
- Songtao Liu and Peng Liu. 2025. High-Layer Attention Pruning with Rescaling. *arXiv preprint arXiv:2507.01900*.
- 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, Online. Association for Computational Linguistics.
- Xinyin Ma, Gongfan Fang, and Xinchao Wang. 2023. [LLM-Pruner: On the Structural Pruning of Large Language Models](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 21702–21720. Curran Associates, Inc.
- Paul Michel, Omer Levy, and Graham Neubig. 2019. Are Sixteen Heads Really Better than One? In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- 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](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.
- Dmitry Molchanov, Arsenii Ashukha, and Dmitry Vetrov. 2017. Variational Dropout Sparsifies Deep Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 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 (CVPR)*, pages 11256–11264.
- John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations*, pages 119–126.
- Leandro Giusti Mugnaini, Bruno Lopes Yamamoto, Lucas Lauton de Alcantara, Victor Zacarias, Edson Bollis, Lucas Pellicer, Anna Helena Real Costa, and Artur Jordao. 2025. Efficient LLMs with AMP: Attention Heads and MLP Pruning. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.
- 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](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 41076–41102. Curran Associates, Inc.
- Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901. Association for Computational Linguistics.
- Xing Niu, Prashant Mathur, Georgiana Dinu, and Yaser Al-Onaizan. 2020. Evaluating Robustness to Input Perturbations for Neural Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8538–8544.
- 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, Online. Association for Computational Linguistics.
- 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 J. Goldsmith, and Mert Pilanci. 2024. [Compressing Large Language Models using Low Rank and Low Precision Decomposition](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 88981–89018. Curran Associates, Inc.
- 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, Cambridge.
- Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. [Patient Knowledge Distillation for BERT Model Compression](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4323–4332, Hong Kong, China. Association for Computational Linguistics.
- Siqi Sun, Zhe Gan, Yu Cheng, Yuwei Fang, 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, Online. Association for Computational Linguistics.
- 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*.

- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal 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. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Elena Voita, David Talbot, Fedor Moiseev, Rico Senrich, and Ivan Titov. 2019. [Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. 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, Brussels, Belgium. Association for Computational Linguistics.
- Boshi Wang, Xiang Yue, Yu Su, and Huan Sun. 2024. [Grokking of Implicit Reasoning in Transformers: A Mechanistic Journey to the Edge of Generalization](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 95238–95265. Curran Associates, Inc.
- Chaoqi Wang, Roger Grosse, Sanja Fidler, and Guodong Zhang. 2019. EigenDamage: Structured Pruning in the Kronecker-Factored Eigenbasis. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6566–6575. PMLR.
- Jiawei Wang, Jiakai Liu, Yuqian Fu, Yingru Li, Xintao Wang, Yuan Lin, Yu Yue, Lin Zhang, Yang Wang, and Ke Wang. 2026a. [Harnessing Uncertainty: Entropy-Modulated Policy Gradients for Long-Horizon LLM Agents](#). In *Forty-third International Conference on Machine Learning*. ArXiv:2509.09265.
- Ziyan Wang, Enmao Diao, Qi Le, Pu Wang, Minwoo Lee, Shu-ping Yeh, Evgeny Stupachenko, Hao Feng, and Li Yang. 2026b. From Local to Global: Revisiting Structured Pruning Paradigms for Large Language Models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 35720–35739.
- Jeffrey T. H. Wong, Cheng Zhang, Xinye Cao, Pedro Gimenes, Christos-Savvas Bouganis, 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](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1513–1528, Dublin, Ireland. Association for Computational Linguistics.
- 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 En-Hsu Yen, Jinxi Zhao, and Zhibin Xiao. 2021. [Rethinking Network Pruning – under the Pre-train and Fine-tune Paradigm](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2376–2382, Online. Association for Computational Linguistics.
- 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*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024a. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.
- 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. 2024b. [LaCo: Large Language Model Pruning via Layer Collapse](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6401–6417, Miami, Florida, USA. Association for Computational Linguistics.
- 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](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 27168–27183. Curran Associates, Inc.
- Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient Surgery for Multi-Task Learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 5824–5836. Curran Associates, Inc.

- Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024. MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 57730–57754. PMLR.
- 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, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 93–104.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. [HellaSwag: Can a Machine Really Finish Your Sentence?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.
- 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 *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 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. 17
B. Proofs and Details	p. 18
B.1. Loss-increase Control via Head Importance (Section 4.2.1)	p. 18
B.1.1. Loss Bound with Operator-Norm Control	p. 18
B.1.2. Norms and Inner Products	p. 18
B.1.3. Estimating $\ W^O\ _2$ and Blockwise Norms via Power Iteration	p. 19
B.1.4. Cross-Entropy Curvature and Propagation to $\mathbf{y}$	p. 19
B.1.5. Why the Quadratic Term Is Typically Negligible	p. 20
B.1.6. Remarks on HIS with Absolute Values	p. 20
B.2. Generalization Gap and Attention Entropy (Section 4.2.2)	p. 20
B.2.1. Notation and Token Averaging	p. 20
B.2.2. Neighboring Datasets and Why They Appear	p. 21
B.2.3. From Attention Perturbation to Output Perturbation	p. 21
B.2.4. Entropy–Total Variation (TV) Control	p. 21
B.2.5. Stability and Generalization	p. 22
B.2.6. Constants and Practical Remarks	p. 22
B.3. Selection Surrogate and HIES Minimization (Section 4.2.3)	p. 22
B.4. Orthogonality and Complementarity (Section 4.2.4)	p. 22
C. Experimental Setup	p. 24
C.1. Experimental Setup for Motivation Study	p. 24
C.2. Model and Dataset	p. 24
C.3. Computing Resources	p. 24
C.4. Dataset Statistics	p. 24
C.4.1. Natural Language Understanding Task	p. 24
C.4.2. Image Classification Task	p. 24
C.4.3. Reasoning Task	p. 24
C.4.4. Multi-modal Vision-Language Task	p. 24
D. Additional Experimental Results	p. 26
D.1. Orthogonality Analysis	p. 26
D.2. Seed Sensitivity	p. 27
D.3. Effect of Calibration Dataset Size	p. 27
D.4. Heatmap of Importance Scores and Pruning Results	p. 27
D.5. 3D Analysis of Attention Head Importance Scores	p. 27
D.5.1. Difference in Pruning Patterns	p. 27
D.5.2. Performance Inversion Across Sparsity Regimes	p. 27
D.6. Experimental Results on Large-scale Reasoning Tasks	p. 27
D.7. Experimental Results on Downstream Tasks	p. 28
D.8. Sensitivity Analysis - Ablation on α	p. 28
D.9. Evaluation Results under Input Perturbations	p. 28
D.10. Evaluation with a Single Fixed α Across Model Families	p. 28
D.11. Combining Attention Entropy with Other Importance Signals	p. 28
E. Efficiency Analysis	p. 29
E.1. Computational Efficiency and FLOPs Reduction	p. 29
E.2. Runtime Overhead and Implementation Details	p. 29
E.3. End-to-End Inference Latency	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., 2020), 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; Jara-dat 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., 2026a). 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 regular-

ization 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 σ Reparam 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 the linear selection surrogate whose exact minimization yields the proposed HIES top- k rule, and we make explicit that this surrogate is a heuristic scalarization rather than an upper bound on the risk (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) loss-control guarantees under operator-norm curvature, (ii) an entropy-based route from stability to generalization, and (iii) a unified, risk-motivated (though heuristic) selection objective 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 $\beta_y := \|\nabla_y^2 \mathcal{L}\|_2$ for the operator-norm curvature at the representation $\mathbf{y}$. Consider the second-order Taylor expansion in $\mathbf{y}$:

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

Taking expectations and using the operator-norm bound yields

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

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

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

hence

$$\Delta \mathcal{L} \leq \sum_{h=1}^{|H|} (1 - m_h) \text{HIS}_h + \frac{\beta_y}{2} \sum_{h=1}^{|H|} (1 - m_h) \|A_h\|_F^2. \quad (10)$$

On the curvature constant β_y . $\beta_y := \|\nabla_y^2 \mathcal{L}\|_2$ is the curvature at the pre-projection representation $\mathbf{y}$, which for a Transformer encoder is an intermediate hidden state rather than the classifier logits: between $\mathbf{y}$ and the logits sit the residual connection, LayerNorm, the feed-forward block, the remaining layers, and the classification head. We therefore treat β_y as a constant of the trained network and estimate it numerically, rather than bounding it in closed form. The familiar cross-entropy curvature bounds ($\frac{1}{4}$ for binary, $\frac{1}{2}$ for multiclass softmax; Appendix B.1.4) hold *at the logits* and are reported there for reference only; none of our reported results depends on a particular numeric value of β_y , since the head ranking is driven by the first-order term.

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 $\Sigma = \text{diag}(\sigma_1, \sigma_2, \dots)$ with $\sigma_1 \geq \sigma_2 \geq \dots \geq 0$, 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

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

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 $W_{(:, \mathcal{I}_h)}^O$, forming

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

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, $W_{(:, \mathcal{I}_h)}^O \in \mathbb{R}^{d \times d_v}$ 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 $\mathbf{y}$

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

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

hence $\|\nabla_z^2 \mathcal{L}\|_2 \leq \frac{1}{4}$.

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

$$\nabla_z^2 \mathcal{L} = \text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top, \quad \|\nabla_z^2 \mathcal{L}\|_2 \leq \frac{1}{2}.$$

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

Scope of these constants. The bounds above are properties of the *classifier logits* $\mathbf{z}$. They do not transfer to $\beta_y := \|\nabla_y^2 \mathcal{L}\|_2$ by composing with the attention output projection alone: $\mathbf{y}W^O$ is a hidden state of width d (e.g., $d = 768$ for BERT_{base}), not a logit vector, and the map from it to the logits traverses the residual connection, LayerNorm, the feed-forward block, the remaining layers, and the classification head. Accordingly we keep β_y as an explicit constant of the trained network (Appendix B.1.1) and do not claim a closed-form bound of the form $c \|W^O\|_2^2$.

B.1.5 Why the Quadratic Term Is Typically Negligible

We first note

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

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,

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

e.g., define

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

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

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

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

The ratio is thus controlled by the curvature β_y together with the activation scale and the alignment floor g :

$$\frac{\text{quadratic}}{\text{first-order}} \leq \frac{\beta_y}{2} \cdot \frac{\max_h \mathbb{E}[\|A_h\|_F]}{g}.$$

A blockwise refinement further tightens this by replacing $\|W^O\|_2^2$ with $\max_h \|W_{(:, I_h)}^O\|_2^2$. 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 (3) is part of the definition to prevent cancellation across samples; consequently, the triangle inequality turns the first-order term into an additive upper bound $\sum_h (1 - m_h) \text{HIS}_h$ (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 $t \in \{1, \dots, n(x)\}$, let $\alpha_t^{(h)}(x) \in \Delta^{n(x)-1}$ 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

$$\begin{aligned} \text{AE}_h(x) &:= \frac{1}{n(x) \log n(x)} \sum_{t=1}^{n(x)} H(\alpha_t^{(h)}(x)) \in [0, 1], \\ \text{AD}_h(x) &:= \frac{1}{n(x) \log n(x)} \sum_{t=1}^{n(x)} \left(\log n(x) - H(\alpha_t^{(h)}(x)) \right) = 1 - \text{AE}_h(x) \in [0, 1]. \end{aligned}$$

For neighboring datasets (S, S') , write the symmetric aggregation

$$\overline{\text{AD}}_h(x) := \frac{1}{2} (\text{AD}_h(x) + \text{AD}'_h(x)).$$

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

B.2.2 Neighboring Datasets and Why They Appear

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

Why neighboring datasets?

- **Symmetrization.** Introduce an i.i.d. ghost sample $S' \sim \mathcal{D}^N$ to rewrite the expected generalization gap as an average of sample-wise differences, e.g., $\mathbb{E}_{S, S'} \left[\frac{1}{N} \sum_{i=1}^N (\ell(f_S; z_i) - \ell(f_{S'}; z'_i)) \right]$, 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 $\mathbb{E}_S [\mathcal{G}(S)] \leq 2\gamma + \frac{B}{N}$.
- **Symmetric inequalities.** Our entropy–total variation (TV) control is symmetric in two distributions (α, α') ; we thus aggregate via $\overline{\text{AD}}_h(x) := \frac{1}{2}(\text{AD}_h(x) + \text{AD}'_h(x))$, 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 $a_h(t) = (\alpha_t^{(h)})^\top V_h \in \mathbb{R}^{d_v}$, hence for neighboring datasets,

$$\Delta_h(t) := a_h(t) - a'_h(t) = (\alpha_t^{(h)} - \alpha_t'^{(h)})^\top V_h.$$

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

$$\|\Delta_h(t)\|_2 \leq \|\alpha_t^{(h)} - \alpha_t'^{(h)}\|_1 \cdot \|V_h\|_{\infty \rightarrow 2} \leq M \|\alpha_t^{(h)} - \alpha_t'^{(h)}\|_1. \quad (11)$$

Averaging over tokens and applying the mask m_h ,

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

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, $\|\mathbf{p} - \mathbf{q}\|_1^2 \leq 4[H(\mathbf{u}) - H(\mathbf{p}) + H(\mathbf{u}) - H(\mathbf{q})]$.*

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

Applying Lemma B.1 to (11) token-wise and averaging,

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

Therefore,

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

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

$$\|\Delta(x)\|_2 \leq \underbrace{\sqrt{8} M \sqrt{|H|(1 - \rho) \log n(x)}}_{=: C_{\text{AE}}(x)} \cdot \sqrt{\sum_{h=1}^{|H|} m_h \overline{\text{AD}}_h(x)}. \quad (12)$$

B.2.5 Stability and Generalization

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

$$\mathbb{E}_S[\mathcal{G}(S)] \leq 2\gamma, \quad \mathcal{G}(S) \leq 2\gamma + \frac{B}{N}.$$

Using (12) and Jensen for $\sqrt{\cdot}$,

$$\gamma \leq L_\ell \mathbb{E}_S[C_{\text{AE}}(x)] \cdot \sqrt{\sum_{h=1}^{|H|} m_h \mathbb{E}_S[\overline{\text{AD}}_h(x)]}.$$

Taking a representative n (e.g., average/max effective length) yields the main-text constant $C_{\text{AE}} = \sqrt{8} M \sqrt{|H|(1-\rho) \log n}$ and Eq. (6).

B.2.6 Constants and Practical Remarks

- **Operator norm.** $\|V_h\|_{\infty \rightarrow 2} := \max_j \|V_h(j, :)\|_2$; take $M := \max_h \|V_h\|_{\infty \rightarrow 2}$ (controlled by Layer-Norm/weight norms).
- **Sequence length.** For padding/causal masking, replace $n(x)$ by the effective context length; averages exclude padded positions.
- **Deficit aggregation.** On-average: $\overline{\text{AD}}_h = \frac{1}{2}(\text{AD}_h + \text{AD}'_h)$; Uniform: $\overline{\text{AD}}_h = \max\{\text{AD}_h, \text{AD}'_h\}$.
- **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 Selection Surrogate and HIES Minimization (Section 4.2.3)

Proof. Let $\text{supp}(m) := \{h : m_h = 1\}$ denote the set of retained heads. Suppose an admissible mask m' with $|\text{supp}(m')| = k$ is not optimal. Then there exist $i \in \text{supp}(m')$ and $j \notin \text{supp}(m')$ such that $\text{HIES}_j > \text{HIES}_i$. Consider the mask $\tilde{m}$ that swaps i and j (retain j , prune i); the constraint in (8) is preserved. The objective in (7) changes by

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

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

B.4 Orthogonality and Complementarity (Section 4.2.4)

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

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

Define

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

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

$$\nabla_{\alpha^{(h)}} \text{HIS}_h = \text{sign}(\alpha^{(h)\top} g_h) g_h, \quad \nabla_{\alpha^{(h)}} \text{AE}_h = -(\mathbf{1} + \log \alpha^{(h)}),$$

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

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

$$\tilde{\nabla} \text{HIS}_h := P \nabla \text{HIS}_h, \quad \tilde{\nabla} \text{AE}_h := P \nabla \text{AE}_h.$$

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

$$\tilde{\nabla}_{\alpha^{(h)}} \text{HIS}_h = P \nabla_{\alpha^{(h)}} \text{HIS}_h = \tilde{u}_h, \quad \tilde{\nabla}_{\alpha^{(h)}} \text{AE}_h = P \nabla_{\alpha^{(h)}} \text{AE}_h = -\tilde{v}_h.$$

Hence

$$\mathbb{E}_S[\langle \tilde{\nabla} \text{HIS}_h, \tilde{\nabla} \text{AE}_h \rangle] = \mathbb{E}_S[\langle \tilde{u}_h, -\tilde{v}_h \rangle] = -\text{tr}(\mathbb{E}_S[\tilde{u}_h \tilde{v}_h^\top]).$$

Decomposing the second moment,

$$\mathbb{E}_S[\tilde{u}_h \tilde{v}_h^\top] = \text{Cov}(\tilde{u}_h, \tilde{v}_h) + \mathbb{E}_S[\tilde{u}_h] \mathbb{E}_S[\tilde{v}_h]^\top.$$

Under $\text{Cov}(\tilde{u}_h, \tilde{v}_h) = 0$ and $\langle \mathbb{E}_S[\tilde{u}_h], \mathbb{E}_S[\tilde{v}_h] \rangle = 0$ (or the stronger $\mathbb{E}_S[\tilde{u}_h] = 0$), we obtain

$$\mathbb{E}_S[\langle \tilde{\nabla} \text{HIS}_h, \tilde{\nabla} \text{AE}_h \rangle] = 0.$$

□

Technical remarks. (i) At $\alpha^{(h)\top} g_h = 0$, use any subgradient of $|\cdot|$ for $\nabla_{\alpha^{(h)}} \text{HIS}_h$. (ii) Since $\alpha^{(h)} = \text{softmax}(\cdot)$, we have $\alpha_j^{(h)} > 0$, so $\log \alpha^{(h)}$ (elementwise) is well-defined. (iii) “ $\text{Cov}(x, y) = 0$ ” 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 $\text{Cov}(u_h, v_h) = 0$ and $\langle \mathbb{E}_S[u_h], \mathbb{E}_S[v_h] \rangle = 0$.

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 BERT_{base} checkpoints that have been fine-tuned by prior work (Devlin et al., 2019), and LLaMA-2_{7B} checkpoint from Hugging Face (Touvron et al., 2023). To examine the generalizability to attention variants and tasks, we further employ ViT_{Large} (Dosovitskiy et al., 2021), LLaVA-1.5_{7B} (Liu et al., 2023), Qwen2.5-3B / 7B (Yang et al., 2024a), 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} (Team Olmo et al., 2025). Table 5 summarizes the parameters and architecture of the four primary models.

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, 2009), Food-101 (Bossard et al., 2014), Fashion MNIST (Xiao et al., 2017), MMLU (Hendrycks et al., 2021), GSM8K (Cobbe et al., 2021), VizWiz-VQA (Gurari et al., 2018), and MM-Vet (Yu et al., 2024).

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 6.

C.4.2 Image Classification Task

Table 6 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., 2021).

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., 2024). The evaluation was conducted using the LLaVA1.5_{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 5: 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-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-1.5 _{7B}	7B	32	32	Multi-modal LLaMA variant integrating a visual encoder; capable of joint image-text understanding and generation.

Table 6: 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.

Table 7: Orthogonality analysis using BERT_{base} on the GLUE benchmark.

Task	$\mathbb{E}[\tilde{u}]$	$\mathbb{E}[\tilde{v}]$	$\text{Cov}(\tilde{u}, \tilde{v})$	$\mathbb{E}[\langle \tilde{u}, \tilde{v} \rangle]$	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 8: Orthogonality analysis using LLaMA-2_{7B} on 5 tasks.

Task	$\mathbb{E}[\tilde{u}]$	$\mathbb{E}[\tilde{v}]$	$\text{Cov}(\tilde{u}, \tilde{v})$	$\mathbb{E}[\langle \tilde{u}, \tilde{v} \rangle]$	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 check of the *uncorrelatedness* hypothesis used in Lemma 4.3. Note that the quantities measured here are the per-head HIS and AE *scores*, not the token-space attention gradients $\tilde{u}_h, \tilde{v}_h$ of the lemma; this check therefore probes the weaker condition $\text{Cov} = 0$ and not the zero-mean condition, which we assume. 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:

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

Consequently,

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

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 stress that this is evidence for $\text{Cov}(\tilde{u}, \tilde{v}) \approx 0$ only; the second hypothesis $\langle \mathbb{E}[\tilde{u}], \mathbb{E}[\tilde{v}] \rangle = 0$ is assumed rather than verified, and Lemma 4.3 should be read as conditional on it.

We further extend the uncorrelatedness analysis across models, with results reported in Table 7 and Table 8. In those two tables the score vectors are normalized but not mean-centered, so the reported $\mathbb{E}[\tilde{u}]$ and $\mathbb{E}[\tilde{v}]$ columns (and hence $\mathbb{E}[\langle \tilde{u}, \tilde{v} \rangle]$, which includes the mean-product term) are not zero by construction; the columns relevant to the hypothesis tested here are Cov and Correlation, whose magnitudes remain small across all tasks and both models.

D.2 Seed Sensitivity

We evaluate the seed sensitivity of HIS and HIES, with results reported in Table 9 and Table 10. For BERT_{base} we report the per-seed standard deviation; for LLaMA-2_{7B} we report the mean over five seeds.

D.3 Effect of Calibration Dataset Size

We analyze the effect of calibration dataset size on the performance and stability of HIES. Tables 11 and 12 report accuracy and standard deviation across calibration sizes ranging from 1 to 1024 samples for BERT_{base} and LLaMA-2_{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

Figure 5 shows the head-importance heatmaps of HIS and HIES on four GLUE tasks (CoLA, MRPC, QNLI, QQP). 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

Figure 6 presents the head-importance scores as 3D surfaces over the layer-head grid, which makes their layer-wise distribution easier to compare.

D.5.1 Difference in Pruning Patterns

As shown in Figure 7, 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 and HIES perform comparably under light pruning: HIS retains a small stability edge and is marginally more accurate on some individual tasks, while the task average slightly favors HIES.

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 and HIES are comparable (HIS holds a small stability edge).
- **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 8 reports reasoning performance on GSM8K and MMLU with LLaMA-2_{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 9 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%. 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,

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

As expected, larger α upweights HIS and preserves heads with strong task relevance, whereas smaller α upweights AE and retains low-entropy, focused heads. For the main results we select $\alpha^\star$ *per task* on a held-out validation split and keep it fixed across pruning ratios; the resulting accuracy–sparsity profiles are shown in Figure 10.

D.9 Evaluation Results under Input Perturbations

We add robustness probes at matched pruning ratios, spanning character/word-level perturbations (Morris et al., 2020; Niu et al., 2020) (Tables 13, 14) and out-of-distribution (OOD) target tasks (ANLI (Nie et al., 2020), SWAG (Zellers et al., 2018), TruthfulQA (Lin et al., 2022)) in Table 15. Under noisy input, our Head Importance-Entropy Score (HIES) improves average accuracy over the Head Importance Score (HIS) by +1.6/+3.2 points at 30%/50% pruning, with comparable or better prediction stability (Table 13). Under controlled perturbation on the larger LLaMA-2-70B, HIES consistently outperforms HIS in both average accuracy (+0.9/+2.9 points) and prediction stability (+1.4/+1.7 points)

at 30%/50% pruning (Table 14). HIES also achieves higher accuracy (+2.0/+2.0 points) and prediction stability (+10.1/+3.5 points) than HIS on the OOD tasks at 30%/50% pruning (Table 15). These results demonstrate that HIES improves both accuracy and prediction stability even under noisy inputs, controlled perturbations, and out-of-distribution settings.

D.10 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 (Yang et al., 2024a), 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} (Team Olmo et al., 2025). In GQA, each query head still produces its own attention distribution over keys, so $\alpha_t^{(h)}$ 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. Table 16 reports Qwen2.5-7B and Qwen3-8B, and Table 17 reports the remaining five models.

D.11 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 11 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 $\approx 4\%$ reduction in FLOPs, while pruning 50% of heads achieves an $\approx 20\%$ reduction. Under HIS-only pruning, model accuracy on TinyBERT with SST-2 sharply degrades beyond a 42% pruning ratio, at which point only an $\approx 16\%$ FLOPs reduction can be attained without critical performance loss. In contrast, HIES-based pruning maintains at least 80% validation accuracy up to a 60% pruning ratio, corresponding to an $\approx 23\%$ 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. This scoring cost is incurred once, before pruning, and is negligible relative to methods that additionally require a recovery fine-tuning pass (e.g., SliceGPT with tuning), which HIES does not use. End-to-end wall-clock latency of the pruned model is reported in Appendix E.3.

E.3 End-to-End Inference Latency

The FLOPs accounting in Appendix E.1 is hardware-independent, so we additionally measure end-to-end wall-clock latency on an NVIDIA H100 GPU with 80 GB of memory (Appendix C.3), the same hardware used for our LLaMA experiments, in fp16. Table 18 reports inference latency of LLaMA-2_{7B} at 30% and 50% head pruning for

three input lengths. Head pruning yields a 1.14–1.26 $\times$ speedup at 1024 tokens, widening to 1.23–1.42 $\times$ at 4096 tokens. The attention computation that pruning removes grows quadratically with sequence length, whereas the FLOPs reduction of Appendix E.1 is dominated by the unpruned feed-forward layers and is therefore length-independent; long-context serving is thus the regime in which head pruning pays off most.

These latency gains follow from the head-pruning ratio itself and are shared by any head-pruning criterion at a matched ratio. What HIES contributes is sustaining accuracy and prediction stability at the aggressive ratios where the gains are realized (Appendix D.10), so the measured speedups are attainable rather than nominal.

Table 9: Accuracy and standard deviation across 5 random seeds for BERT_{base} on the GLUE benchmark. STS-B is reported here as the raw Pearson correlation ($\times 100$), whereas Table 1 rescales it to $[0, 100]$; that column is therefore not directly comparable between the two tables. CoLA is on the same rescaled $[0, 100]$ axis in both tables.

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

Table 10: Mean accuracy across 5 random seeds for LLaMA-2_{7B} on HellaSwag, Winogrande, and ARC-c. Relative improvements of HIES over HIS are shown in blue.

Pruning Ratio (%)		HIS Accuracy (%)	HIES (ours) Accuracy (%)
HellaSwag	10	53.63	53.98 0.66%↑
	20	51.70	52.36 1.28%↑
	30	49.00	49.84 1.71%↑
	40	45.27	46.27 2.20%↑
	50	37.01	36.13 2.37%↓
Winogrande	10	66.38	67.01 0.95%↑
	20	64.07	66.27 3.42%↑
	30	63.03	64.67 2.60%↑
	40	60.16	61.06 1.50%↑
	50	56.16	55.83 0.59%↓
ARC-c	10	33.15	35.05 5.73%↑
	20	33.15	36.27 9.41%↑
	30	32.34	35.46 9.64%↑
	40	29.69	34.24 15.30%↑
	50	25.29	28.95 14.48%↑
Avg.	10	51.05	52.01 1.88%↑
	20	49.64	51.63 4.01%↑
	30	48.12	49.99 3.88%↑
	40	45.04	47.19 4.76%↑
	50	39.49	40.30 2.07%↑

Table 11: 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

Table 12: 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 13: Robustness under noisy inputs on LLaMA-2_{7B}. Accuracy and prediction stability (%) are reported under two pruning ratios.

Prune%	Method	Accuracy (%)						Stability (%)					
		ARC-C	ARC-E	PIQA	OBQA	WinoG	Avg	ARC-C	ARC-E	PIQA	OBQA	WinoG	Avg
30%	HIS	31.5	62.7	74.0	21.0	63.0	50.4	81.7	83.7	93.7	81.7	83.3	84.8
	HIES (ours)	33.2	62.3	75.3	23.0	66.0	52.0	79.3	86.3	93.0	79.3	81.7	83.9
50%	HIS	23.7	49.7	63.7	14.0	56.0	41.4	61.7	65.0	78.7	65.3	66.3	67.4
	HIES (ours)	27.1	53.3	65.7	17.0	59.7	44.6	64.4	66.0	80.7	64.3	70.0	69.1

Table 14: Robustness under input perturbations on LLaMA-2_{70B}. Accuracy and prediction stability (%) are reported under two pruning ratios.

Prune%	Method	Accuracy (%)						Stability (%)					
		ARC-C	ARC-E	PIQA	OBQA	WinoG	Avg	ARC-C	ARC-E	PIQA	OBQA	WinoG	Avg
30%	HIS	35.6	72.0	76.0	24.7	72.7	56.2	83.1	90.3	93.3	82.0	89.3	87.6
	HIES (ours)	38.0	70.7	76.7	26.3	73.7	57.1	85.4	90.7	94.7	84.3	89.7	89.0
50%	HIS	29.2	60.7	73.3	20.7	62.7	49.3	73.2	77.3	88.0	76.0	73.3	77.6
	HIES (ours)	33.6	66.3	74.3	22.7	64.3	52.2	76.9	79.0	89.7	80.7	70.3	79.3

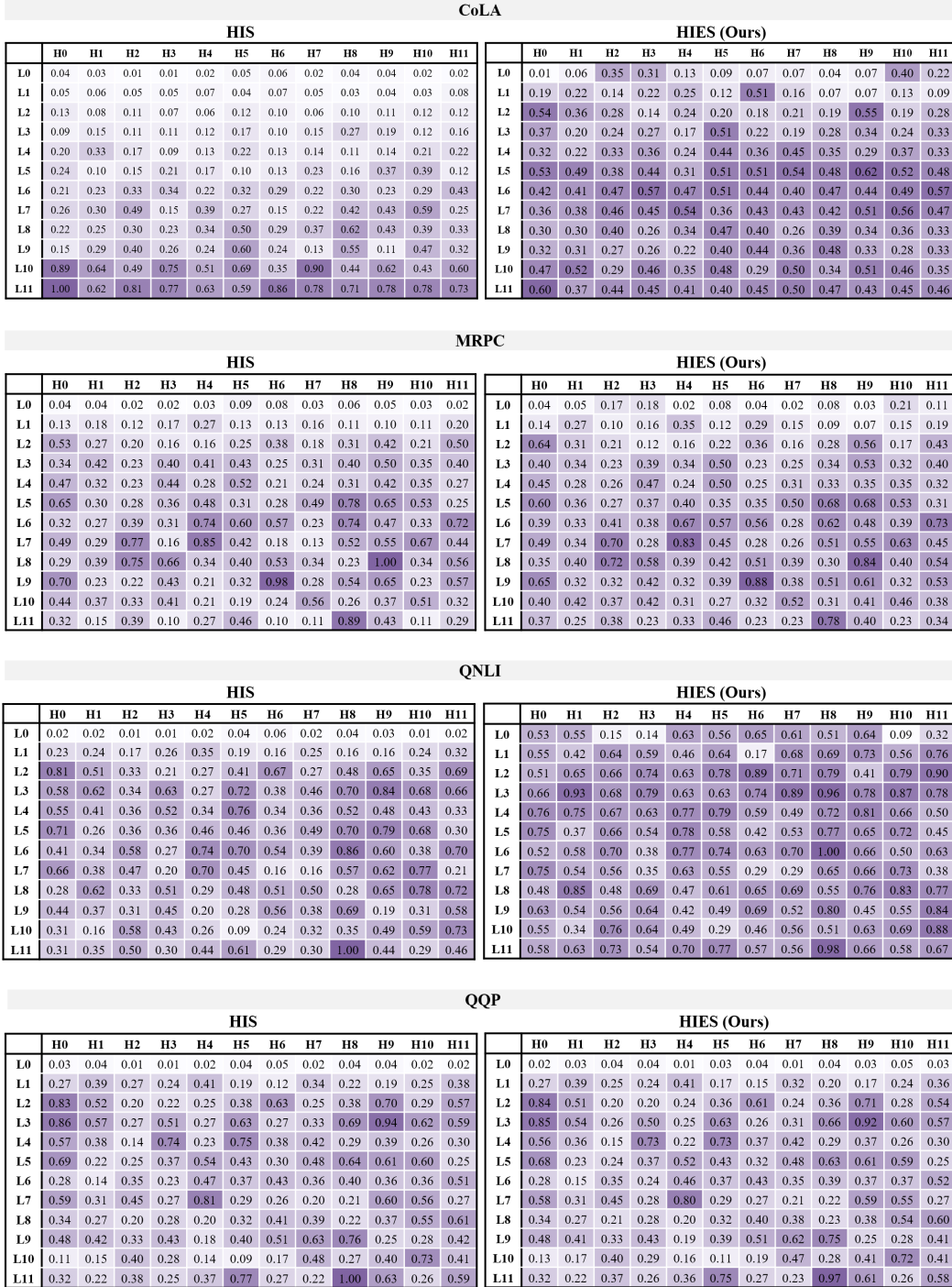
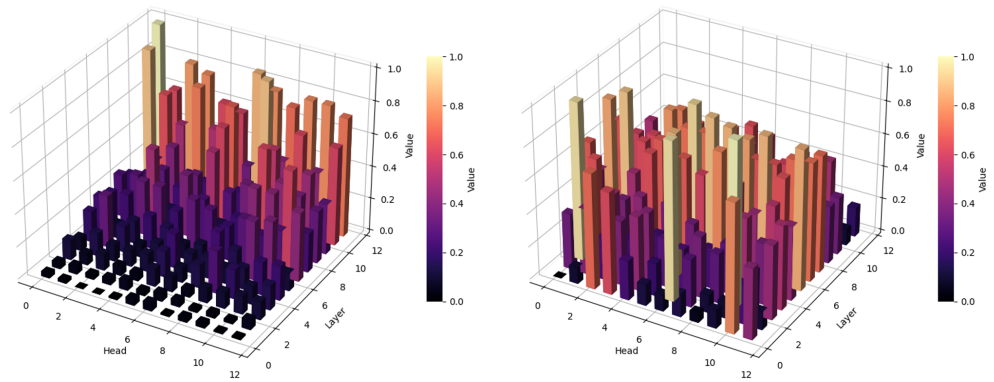
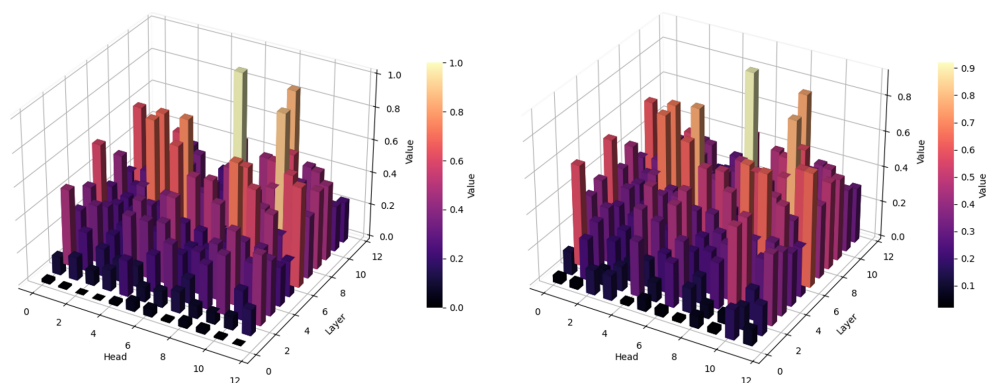


Figure 5: Heatmaps of head-importance scores across four GLUE tasks (CoLA, MRPC, QNLI, QQP). Left: HIS; Right: HIES (ours). Rows = layers (L0–L11); columns = heads (H0–H11).

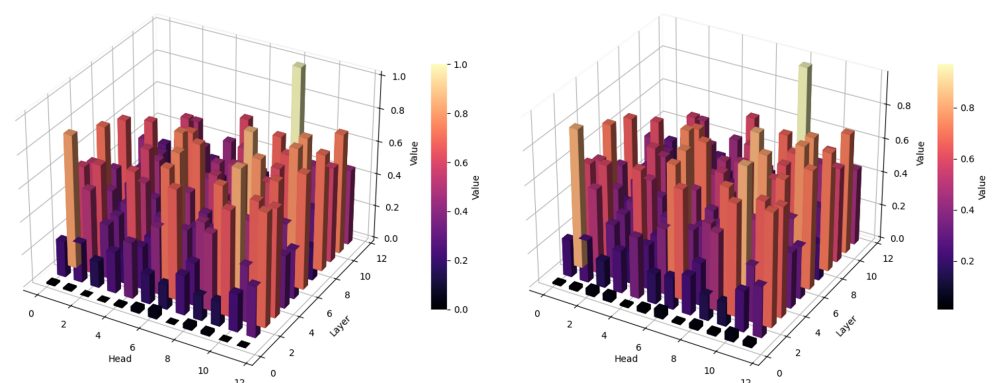
CoLA



MRPC



QNLI



WNLI

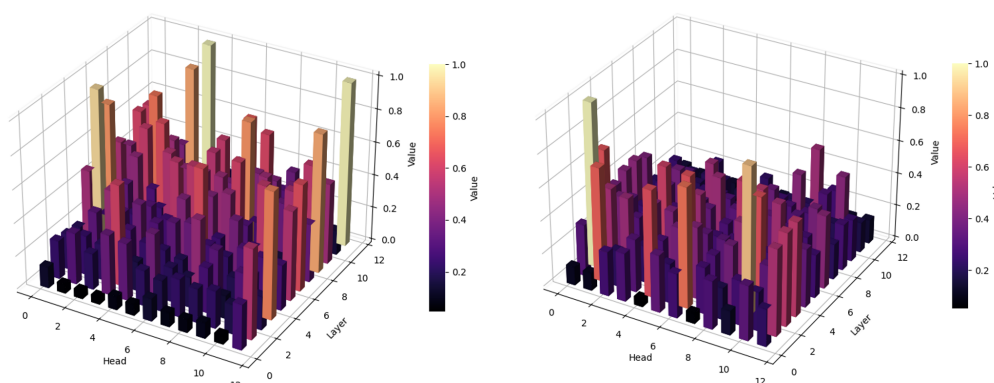


Figure 6: 3D Analysis of Attention Head Importance Scores

CoLA

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	0.44
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	0.01	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	0.44
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	0.01	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	0.44
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	0.01	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

46930

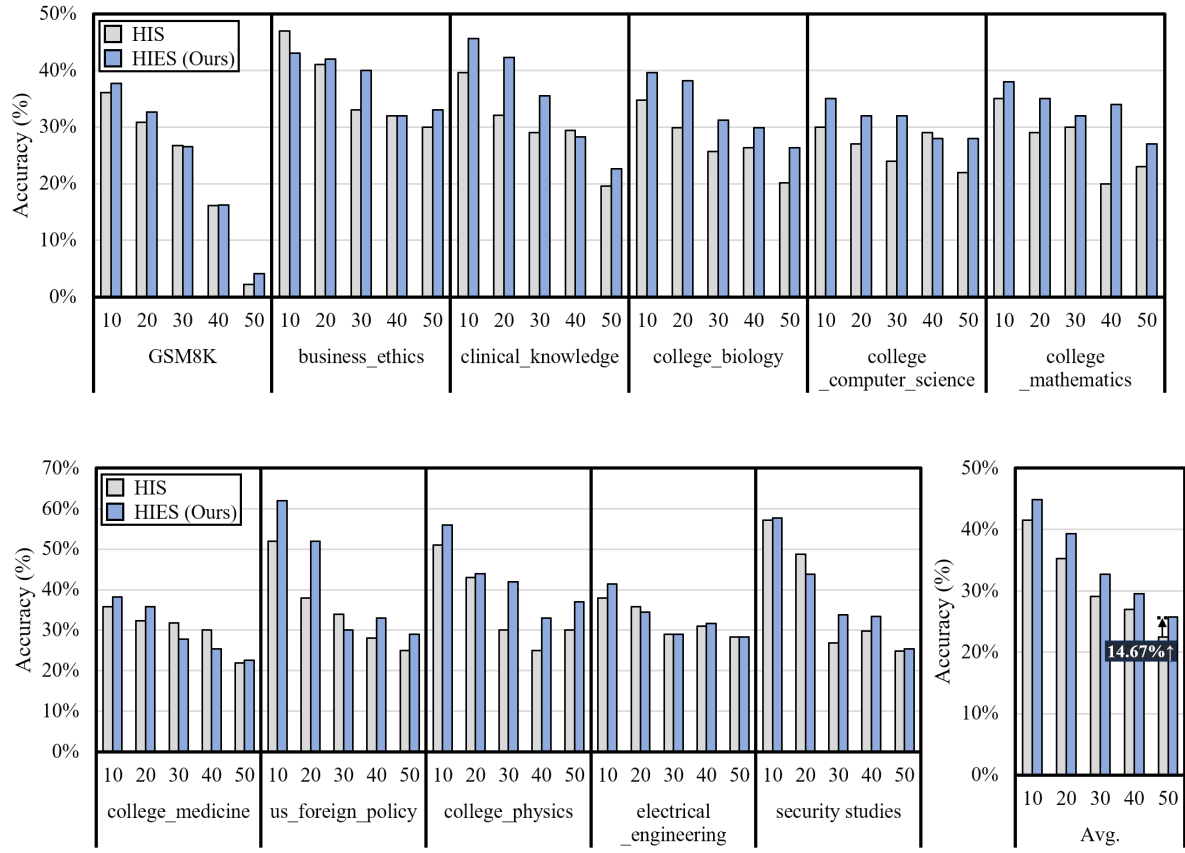


Figure 8: Results on GSM8K (math word-problem reasoning) and MMLU (10-task knowledge reasoning) with LLaMA-2_{7B}.

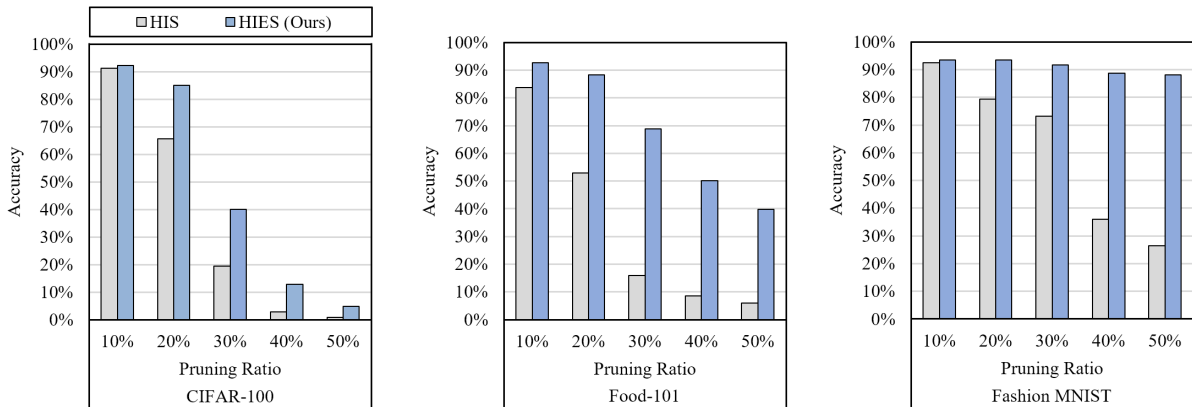


Figure 9: Evaluation of HIES on the image classification benchmarks. HIES consistently outperforms baseline, demonstrating robust and stable performance across downstream tasks.

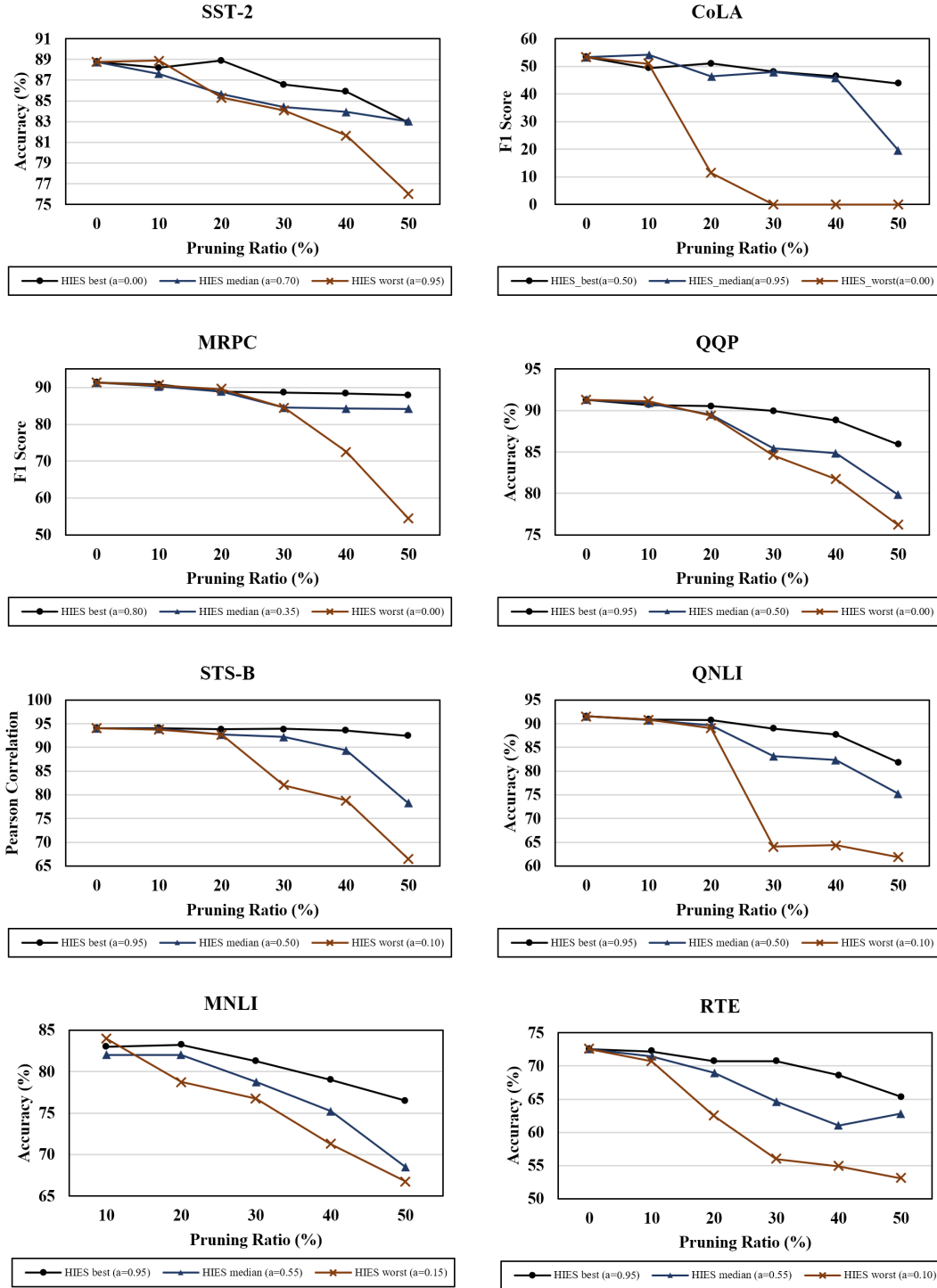


Figure 10: HIES sensitivity to the mixing coefficient α on GLUE. For each task, we sweep α and report three choices— α_{best} , α_{median} , α_{worst} —selected by weighted AUC (wAUC) across pruning ratios. Curves plot performance versus pruning ratio for these three settings.

Table 15: Robustness on out-of-distribution target tasks using LLaMA-2_{7B}. Accuracy and prediction stability (%) are reported under two pruning ratios.

Target Task	Prune%	Method	Accuracy (%)						Stability (%)					
			ARC-C	ARC-E	PIQA	OBQA	WinoG	Avg	ARC-C	ARC-E	PIQA	OBQA	WinoG	Avg
ANLI-R1	30%	HIS	26.0	26.3	27.7	26.3	26.3	26.5	36.3	37.3	42.0	38.0	36.3	38.0
		HIES (ours)	36.7	32.3	32.0	29.0	35.0	33.0	58.3	54.3	59.3	45.7	56.3	54.8
	50%	HIS	26.0	26.0	26.0	26.0	25.7	25.9	36.3	37.3	36.3	36.3	36.3	36.5
		HIES (ours)	26.3	26.3	30.7	31.7	27.7	28.5	39.0	40.0	47.3	42.3	43.3	42.4
ANLI-R2	30%	HIS	34.0	34.0	33.0	34.3	33.3	33.7	42.3	42.0	47.0	42.0	43.7	43.4
		HIES (ours)	35.3	36.7	32.7	33.3	33.0	34.2	60.7	56.0	56.0	51.0	54.3	55.6
	50%	HIS	34.0	33.7	34.3	34.3	34.0	34.1	42.3	42.7	42.7	42.3	42.0	42.4
		HIES (ours)	36.3	35.7	38.3	37.0	36.3	36.7	43.0	44.3	46.0	42.3	46.0	44.3
ANLI-R3	30%	HIS	36.3	37.0	34.7	37.0	37.7	36.5	39.0	41.0	45.7	42.0	40.3	41.6
		HIES (ours)	37.7	40.7	37.7	39.3	40.3	39.1	62.7	62.3	59.3	59.0	60.7	60.8
	50%	HIS	37.0	36.3	37.0	36.7	36.7	36.7	38.0	37.3	37.0	36.7	37.0	37.2
		HIES (ours)	38.0	38.3	35.3	33.7	39.0	36.9	36.3	40.0	39.3	43.7	42.3	40.3
SWAG	30%	HIS	53.7	53.0	53.0	53.7	52.7	53.2	91.7	90.3	91.3	92.7	92.3	91.7
		HIES (ours)	53.7	53.7	54.0	54.3	54.0	53.9	93.3	93.3	93.3	94.7	94.0	93.7
	50%	HIS	42.0	40.0	40.0	41.7	41.3	41.0	78.3	77.0	77.0	77.7	78.0	77.6
		HIES (ours)	45.0	43.7	40.7	45.7	46.7	44.4	84.3	83.0	80.0	84.0	83.7	83.0
TruthfulQA	30%	HIS	22.7	23.3	23.0	24.7	23.7	23.5	82.0	81.0	79.0	78.7	80.7	80.3
		HIES (ours)	22.7	22.3	22.3	22.7	25.0	23.0	81.0	82.0	80.7	79.7	80.3	80.7
	50%	HIS	21.0	21.3	21.0	20.3	21.3	21.0	65.0	65.7	63.3	62.7	65.3	64.4
		HIES (ours)	22.0	21.3	22.3	22.7	23.3	22.3	65.3	68.3	61.0	65.3	68.3	65.6

Table 16: 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 ($\alpha=0.9$ for Qwen2.5-7B, $\alpha=0.7$ 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, $\alpha=0.9$)						
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, $\alpha=0.7$)						
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

Table 17: 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 five models.

Pruning	Method	HellaSwag	ARC-C	ARC-E	Winogrande	Avg.
Llama-3.1-8B (GQA, 32Q/8KV, $\alpha=0.1$)						
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, $\alpha=0.1$)						
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, $\alpha=0.6$)						
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, $\alpha=0.9$)						
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, $\alpha=0.4$)						
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

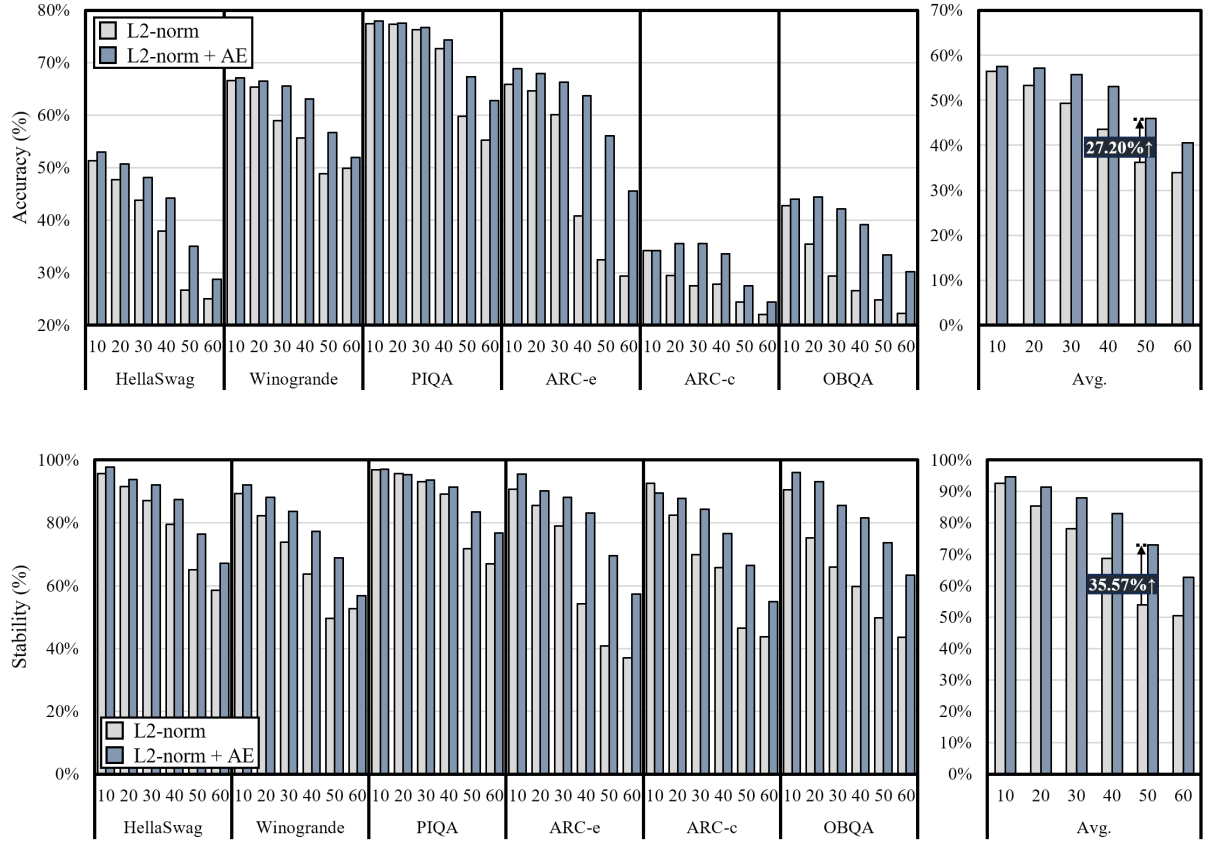


Figure 11: Accuracy and stability improvements when combining AE with gradient L2-norm–based importance scores, compared to using the L2-norm alone. We conduct experiments on LLaMA-2_{7B} for HellaSwag, Winogrande, PIQA, ARC-e, ARC-c, and OBQA.

Table 18: End-to-end inference latency of LLaMA-2_{7B} under head pruning (fp16), measured on the H100 GPU described in Appendix C.3. Speedup is relative to the unpruned model at the same sequence length.

SeqLen	Pruning	Latency (ms)	Speedup
1024	Dense (0%)	293.57	1.00×
	30%	256.55	1.14×
	50%	232.33	1.26×
2048	Dense (0%)	683.36	1.00×
	30%	570.09	1.20×
	50%	505.20	1.35×
4096	Dense (0%)	1586.07	1.00×
	30%	1287.03	1.23×
	50%	1118.82	1.42×