RAHP: ROBUSTNESS-AWARE HEAD PRUNING FOR CERTIFIED TRANSFORMER MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Transformers lie at the core of modern AI, yet their susceptibility to adversarial perturbations raises reliability concerns. Empirical defenses often lack guarantees, while certification-based approaches provide them at nontrivial computational cost. We introduce RAHP (Robustness-Aware Head Pruning), a certification-guided pruning framework for Transformers. RAHP scores each attention head with a composite of (i) ΔCLEVER, the predicted increase in a certified-robustness lower bound when masking that head, and (ii) Fisher information, the estimated accuracy cost of removing it. We prune heads that maximize robustness gain per accuracy cost. Across evaluated tasks, RAHP yields compact models with stronger CLEVER lower bounds and minimal change in clean accuracy, and it improves resistance to a wide variety of strong attacks. By leveraging a certified metric to steer structural pruning, RAHP makes certification-oriented robustness more practical and scalable.

1 INTRODUCTION

The field of artificial intelligence has witnessed a paradigm shift with the emergence of transformer architectures, which have revolutionized not only natural language processing but also computer vision, speech recognition, and numerous other domains. However, as these models become increasingly common in critical applications, a fundamental question emerges: **how robust are these systems when faced with unexpected inputs, or adversarial attacks?**

In recent years, research has revealed critical vulnerabilities in Transformer-based models, demonstrated through concrete adversarial attacks that exploit these weaknesses. These attacks are often invisible to human readers: small changes at the token or character level, harmless word substitutions, or reworded sentences may appear trivial, yet they can cause a model to make drastically different decisions. These manipulations occur at various levels. At the character level, minor edits (e.g., "hotel" $\rightarrow$ "h0tel") preserve readability but can mislead the model (Ebrahimi et al., 2017). At the word level, replacing a word with a semantically similar alternative (e.g., "terrible" $\rightarrow$ "awful") retains the meaning for humans but alters the model's internal representation (Gan et al., 2021). At the sentence level, paraphrasing (e.g., "The movie was surprisingly good." $\rightarrow$ "I was taken aback by how enjoyable the film was.") maintains the intended message but may shift the model's response (Krishna et al., 2023). Such perturbations can lead to changes in embedding vectors, noisy or misleading representations, out-of-vocabulary tokens, or fragmented tokenization, which ultimately result in unstable or incorrect model behavior.

Robustness refers to a model's ability to maintain reliable performance in the face of variations, imperfections, or challenges in the input data (Freiesleben & Grote, 2023). This includes resilience to noise, shifts in data distribution, and deliberate adversarial manipulations (Brown et al., 2023). Broadly speaking, robustness can be categorized into three key types. *Adversarial robustness* concerns the model's ability to resist carefully crafted perturbations designed to cause incorrect outputs, even when the changes are imperceptible to humans (Shao et al., 2021). *Distributional robustness* focuses on how well a model generalizes when the test data distribution deviates from the training distribution, a common challenge in real-world deployment (Samuel & Chechik, 2021). *Certified robustness*, in contrast, offers formal guarantees. It quantifies the maximum perturbation under which a model's prediction is guaranteed to remain unchanged, typically using techniques rooted in provable bounds or integer programming methods (Zeng et al., 2023; Kumar et al., 2023).

While adversarial and distributional robustness are essential in practice, they often rely on empirical testing or assumptions about the nature of future inputs. Certified robustness provides a rigorous, model-agnostic foundation for developing robustness-oriented methods, making it a principled and compelling focus for algorithmic design and research.

Building on these three categories, the mechanisms to achieve robustness differ in emphasis. For adversarial robustness, the dominant recipe is to expose the model to worst-case perturbations during training (adversarial training) and/or add sensitivity-controlling penalties such as input-gradient regularization or virtual adversarial training that smooths the output locally around each sample (Madry et al., 2017; Ross & Doshi-Velez, 2018). For distributional robustness, methods optimize for performance under plausible distribution shifts rather than single samples, e.g., group DRO to raise worst-group accuracy and invariance-seeking objectives such as Invariant Risk Minimization (IRM) for OOD generalization (Sagawa et al., 2019; Arjovsky et al., 2019). For certified robustness, training and evaluation rely on provable bounds that guarantee prediction invariance within a perturbation set, via randomized smoothing, or convex relaxations of the adversarial region (Cohen et al., 2019; Wong & Kolter, 2018); metrics such as CLEVER (Weng et al., 2018) provide attack-agnostic lower bounds that are widely used to assess robustness even when certification is not directly optimized. A key strength of certified approaches is that their guarantees are independent of any particular attack strategy, enabling model and threat-agnostic comparisons across methods.

Pruning removes parameters, neurons, or components (e.g., attention heads) to simplify the network while preserving accuracy. Beyond efficiency, pruning can also shape decision geometry by reducing unstable non-linearities and eliminating fragile pathways, which has been observed to improve empirical robustness, especially when combined with robust training objectives (Sehwag et al., 2020). Recent work further shows that pruning can improve certified robustness: relaxing or removing unstable ReLUs can tighten verification bounds and raise certified accuracy, and strategically grafting linearity in place of weak non-linear units likewise boosts certifiable guarantees (Zhangheng et al., 2022; Chen et al., 2022). In this spirit, we extend robustness-aware pruning to Transformers: we use the CLEVER score (Weng et al., 2018) as a guiding, certification-oriented signal alongside accuracy-preservation criteria to rank attention heads for removal, and show across tasks that targeting pruning by a certified-robustness metric increases a model's robustness while minimally affecting clean performance.

In this work, we translate this certified-robustness-guided pruning idea into practice by introducing **RAHP: Robustness-Aware Head Pruning**. RAHP treats every attention head as a candidate for removal and scores it with two complementary scores: *(i) Fisher information*, which estimates the accuracy loss incurred if the head is pruned, and *(ii)* Δ*CLEVER*, which estimates the increase in certified robustness achieved by masking that head. By pruning the heads that maximize a weighted trade-off between these scores, RAHP compresses the model while widening its provable perturbation radius. Because the CLEVER score is embedded directly in the pruning rule, RAHP is, to our knowledge, the first Transformer pruning framework that optimizes certified robustness and efficiency simultaneously, achieving stronger guarantees without costly adversarial retraining and with negligible impact on clean accuracy.

2 Related Works

Prior work strengthens Transformer robustness by smoothing predictions or explicitly optimizing against perturbations. R-Drop enforces output consistency across different dropout masks via a bidirectional KL term, reducing variance and sharpening decision margins (Wu et al., 2021). Child-Tuning masks gradients to update only a "child" subset of parameters, stabilizing fine-tuning on limited data and improving robustness without overfitting the full model (Xu et al., 2021). SMART augments fine-tuning with small, principled adversarial perturbations plus a KL smoothness regularizer in representation space (Jiang et al., 2019). FreeLB performs multi-step adversarial training in the embedding space to craft stronger perturbations while accumulating informative gradients (Zhu et al., 2019). These methods improve robustness but offer no attack-agnostic guarantees.

A complementary line pursues certified robustness with formal guarantees. Randomized smoothing wraps a base classifier with Gaussian noise to certify an ℓ_2 radius per input (Cohen et al., 2019). Symbolic/interval bound propagation adapts certification to NLP by bounding worst-case effects of discrete edits such as synonym or character substitutions (Huang et al., 2019). Finally, CLEVER

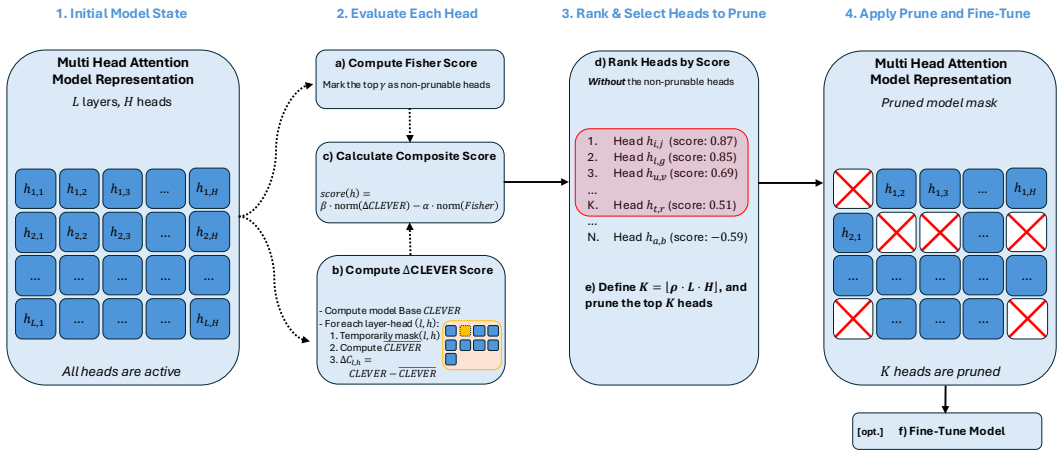

Figure 1: **RAHP Method Overview:** For each head, (a) compute Fisher score, (b) compute ΔCLEVER score, and (c) combine into a composite score. (d) Mark the top-γ heads by Fisher as non-prunable and rank the remaining heads by composite score. (e) Prune the top-K heads.

estimates attack-agnostic lower bounds on the minimal adversarial distortion, and is widely used as a certification-oriented assessment metric even when certificates are not computed (Weng et al., 2018). These approaches clarify trade-offs between guarantees and accuracy/compute.

Pruning has emerged as a structural route to robustness by removing fragile pathways or reducing unstable non-linearities. Beyond empirical gains, recent studies show pruning can improve certified robustness by tightening verification bounds and reducing neuron instability (Zhangheng et al., 2022; Chen et al., 2022). In parallel, ROSE (Robust Selective Fine-Tuning) selectively updates robust parameters during adaptation, improving adversarial robustness and offering a strong baseline for comparison in NLP (Jiang et al., 2022). Our work follows this structural perspective for Transformers, but differs by using a certified-robustness–oriented signal (CLEVER) to guide which attention heads to remove.

In our experiments, we compare against these approaches to evaluate the benefits of our method.

3 Methodology

As illustrated in Figure 1, we propose a two-score pruning framework that, in a single global step across all layers, removes self-attention heads to enhance robustness without sacrificing accuracy.

Concretely, each attention head is assigned two complementary scores: a **CLEVER**-based robustness score that reflects the change in robustness when the head is masked, and a **Fisher**-based importance score that quantifies the accuracy cost of its removal. After normalization, the two signals are combined into a composite score, and all heads are ranked globally. To safeguard accuracy, a fixed fraction γ of the most Fisher-important heads are explicitly protected from pruning. The remaining heads are pruned according to the target pruning ratio ρ, removing those with the highest composite scores. Finally, the pruned model can be optionally fine-tuned slightly on the training data to mitigate potential clean performance loss.

3.1 Notation

We define the notation used throughout this paper as follows. Let $f(\cdot)$ denote a Transformer-based sequence classifier. The model produces logits $z = f(x) \in \mathbb{R}^C$ corresponding to C output classes.

The Transformer model consists of L attention layers, where each layer $\ell \in \{0, \ldots, L-1\}$ contains H attention heads. A specific head is identified by the tuple (ℓ, h), with $h \in \{0, \ldots, H-1\}$. A binary head mask is defined as $M \in \{0,1\}^{L \times H}$, where $M_{\ell,h} = 0$ indicates that the head (ℓ, h) is pruned, and $M_{\ell,h} = 1$ indicates that it is retained. Finally, the classification margin for a given

Algorithm 1 Robustness-Aware Head Pruning

Input: Model f; dataset D; weights α, β; prune ratio ρ; non-prunable fraction γ
Output: Pruned head mask $M \in \{0,1\}^{L \times H}$

Initialize $M \leftarrow \mathbf{1}^{L \times H}$ *(all heads active)*
for $\ell = 0$ **to** $L-1$ **do**
  for $h = 0$ **to** $H-1$ **do**
    $M' \leftarrow M$ with $M'_{\ell,h} \leftarrow 0$
    $\Delta C_{\ell,h} \leftarrow \mathbb{E}_{x \sim D}\left[\frac{g(f(x,M'))}{\|\nabla_x g(f(x,M'))\|_2 + \varepsilon} - \frac{g(f(x,M))}{\|\nabla_x g(f(x,M))\|_2 + \varepsilon}\right]$
    $F_{l,h} \leftarrow \mathbb{E}_{(x,y)\sim D}\left\|\nabla_{\theta_{l,h}} J(f(x), y)\right\|_2^2$
    $I_{\ell,h} \leftarrow \log(F_{\ell,h} + \varepsilon)$
  end for
end for
$I'_{\ell,h} \leftarrow \text{normalize}(I_{\ell,h})$
$\Delta C'_{\ell,h} \leftarrow \text{normalize}(\Delta C_{\ell,h})$
$\mathcal{N} \leftarrow$ top-γ fraction of heads with highest $I_{\ell,h}$ *(non-prunable heads)*
for (ℓ, h) over all heads **do**
  if $(\ell, h) \in \mathcal{N}$ **then**
    $S_{\ell,h} \leftarrow -\infty$ *(protect non-prunable heads)*
  else
    $S_{\ell,h} \leftarrow \beta \cdot \Delta C'_{\ell,h} - \alpha \cdot I'_{\ell,h}$
  end if
end for
$K \leftarrow \lfloor \rho \cdot L \cdot H \rfloor$
Prune top-K heads with largest $S_{\ell,h}$ by setting $M_{\ell,h} \leftarrow 0$
Fine-tune f on D using the updated M *(optionally)*
return M

output is defined as $g(z) = z_c - z_{c'}$, where z_c is the logit of the correct class, and $z_{c'}$ is the highest logit among all incorrect classes.

3.2 CLEVER-Based Robustness Score

To quantify model robustness, we adapt the CLEVER (Cross-Lipschitz Extreme Value for nEtwork Robustness) score (Weng et al., 2018). The CLEVER score provides a lower bound on the minimum perturbation needed to cause a misclassification, with higher scores indicating greater robustness. It is formally defined as the minimum ratio of the classification margin to the norm of its gradient with respect to the input, approximated via extreme value theory.

We estimate this score for a batch of samples, where the score for a single input x is:

$$C(x, M) = \frac{g(f(x, M))}{\|\nabla_x g(f(x, M))\|_2 + \epsilon}. \tag{1}$$

Here, $f(x, M)$ is the model's forward pass using the head mask M, and ϵ is a small constant for numerical stability. The gradient is taken with respect to the input embeddings.

To assess the contribution of an individual head (ℓ, h), we measure the change in CLEVER score when this head is removed. Formally, letting M be the current mask and $M'_{\ell,h}$ the same mask but with head (ℓ, h) pruned, we define the **Robustness Score** as:

$$\Delta C_{\ell,h} = \mathbb{E}_{x \sim D}\left[C(x, M'_{\ell,h}) - C(x, M)\right] \tag{2}$$

where the expectation is taken over the dataset D.

The interpretation of $\Delta C_{\ell,h}$ is straightforward:

- If $\Delta C_{\ell,h} = 0$, pruning head (ℓ, h) has no effect on robustness.
- If $\Delta C_{\ell,h} > 0$, the model becomes more robust after pruning (ℓ, h), making them good pruning candidates.
- If $\Delta C_{\ell,h} < 0$, pruning (ℓ, h) weakens robustness and such heads are best kept.

In practice, the higher the $\Delta C_{\ell,h}$ value, the more robust the pruning of head (ℓ, h) contributes to the overall model. This makes ΔCLEVER a natural robustness-oriented reward signal within our composite pruning criterion.

3.3 Fisher-Based Accuracy Score

To prevent the pruning process from significantly degrading model accuracy, we must quantify the importance of each attention head to the model's primary task. For this, we use the *Fisher Information* (FI), which measures the sensitivity of the model's loss to changes in its parameters. Because FI provides a principled sensitivity estimate, it has been widely adopted as a pruning criterion (Molchanov et al., 2016; Theis et al., 2018; Molchanov et al., 2019; Liu et al., 2021; Kwon et al., 2022; Sung et al., 2021). Heads with high FI are considered more critical to the model's performance, as pruning them would likely incur a large accuracy cost.

We approximate the diagonal of the FI matrix for the parameters $\theta_{\ell,h}$ of each head (ℓ, h). For a single data point (x, y), the FI is estimated as the squared gradient of the loss function J (e.g., Cross-Entropy) with respect to the head's parameters:

$$F_{l,h}(x, y) = \left\| \nabla_{\theta_{l,h}} J(f(x), y) \right\|_2^2 . \tag{3}$$

Averaging over the dataset D yields the head-level Fisher estimate:

$$F_{l,h} = \mathbb{E}_{(x,y)\sim D} \left[F_{l,h}(x, y) \right] . \tag{4}$$

For numerical stability and to temper the heavy-tailed distribution of Fisher values, we apply a log compression before scoring. The resulting $F_{\ell,h}$ acts as the **Accuracy Score**: heads with high Fisher values are strongly tied to minimizing task loss and should therefore be preserved, whereas heads with low Fisher values tend to contribute little to accuracy and are good candidates for pruning.

3.4 Composite Score & Pruning Rule

After computing both the robustness and accuracy metrics for all heads, we combine them into a single signal that guides the pruning decision. Specifically, we normalize each metric across the entire model to a $[0, 1]$ range, obtaining $\Delta C'_{\ell,h}$ (normalized robustness gain) and $I'_{\ell,h}$ (normalized accuracy cost). To stabilize the scale of the Fisher estimates, we first apply a log transformation $I_{\ell,h} = \log(F_{\ell,h} + \varepsilon)$, and then normalize $I_{\ell,h}$ across all heads to yield $I'_{\ell,h}$.

The **Composite Score** $S_{\ell,h}$ for head (ℓ, h) is then defined as:

$$S_{\ell,h} = \beta \cdot \Delta C'_{\ell,h} - \alpha \cdot I'_{\ell,h}, \tag{5}$$

where α and β control the trade-off between robustness and accuracy preservation. A higher value of $S_{\ell,h}$ indicates that pruning the head yields stronger robustness improvements (large $\Delta C'_{\ell,h}$) while incurring only a small accuracy penalty (low $I'_{\ell,h}$).

To ensure that accuracy-critical heads are not mistakenly removed, we protect a fixed fraction γ of heads with the highest Fisher values (the *non-prunable set*). These are excluded from consideration regardless of their composite score. The remaining heads are globally ranked by $S_{\ell,h}$ in descending order, and pruning is applied to the top $K = \lfloor \rho \cdot L \cdot H \rfloor$ heads. Sorting in descending order ensures that we remove precisely those heads with the best trade-off: the highest robustness gain combined with the lowest accuracy cost. Put differently, the higher the composite score, the more it reflects a desirable balance of large robustness gains with minimal accuracy loss.

Finally, the pruned model can be optionally fine-tuned on the original dataset to recover any residual performance loss. This one-shot, global ranking strategy enables RAHP to jointly optimize robustness and accuracy without iterative per-layer pruning. The complete algorithm is summarized in Algorithm 1.

4 EXPERIMENTS

4.1 DATASETS & MODELS

We evaluate our method using four tasks from the GLUE (Wang et al., 2018) and AdvGLUE (Wang et al., 2021) benchmarks. GLUE is a widely used benchmark for natural language understanding that evaluates model performance on clean, human-written text. AdvGLUE extends this by providing human-verified adversarial counterparts for each GLUE task, constructed via 14 diverse attack techniques targeting model vulnerabilities. This allows us to measure robustness against realistic adversarial perturbations. To ensure reliable comparison, we avoid synthetic or automatic attack generation, which often introduces invalid or semantically ambiguous examples.

SST-2 (Socher et al., 2013) is a binary sentiment classification task where models must predict whether a single sentence from a movie review expresses positive or negative sentiment.

RTE (Bentivogli et al., 2009) is a natural language inference task derived from multiple textual entailment challenges. Each example consists of a premise and a hypothesis, and the model must determine whether the hypothesis can be logically inferred from the premise.

QNLI (Rajpurkar et al., 2016) is a question-answering-derived task where the model must decide whether a given context sentence contains the answer to a corresponding question.

QQP (Quora, 2018) is a large-scale semantic similarity benchmark based on real user-submitted questions from Quora. Each pair of questions must be classified as paraphrases or not, testing a model's ability to detect semantic equivalence and redundancy.

For this experiment, we use two widely adopted Transformer architectures: ***RoBERTa$_{BASE}$***, with 125 million parameters, 12 layers, and 12 attention heads per layer; and ***RoBERTa$_{LARGE}$***, with 355 million parameters, 24 layers, and 16 attention heads per layer. These models are strong baselines for both standard and adversarial evaluations.

4.2 EXPERIMENTAL SETTINGS

Across all experiments, we fixed the pruning hyperparameters to values that provided the most stable and effective performance. Specifically, we set the trade-off weights to $\beta = 1$ and $\alpha = 0.5$, which we found to be the most effective combination for balancing robustness and importance. You can find a detailed analysis of alternative weightings in Section 5.

For the pruning configuration, we employed a prune ratio of 60%, meaning that only 40% of the attention heads remain active after pruning. In addition, we enforced a non-prunable fraction of $\gamma = 10\%$. Here as well, a detailed analysis can be found in Section 5.

Model	Pruning Volume	SST-2 GLUE	SST-2 AdvGLUE	RTE GLUE	RTE AdvGLUE	QNLI GLUE	QNLI AdvGLUE	QQP GLUE	QQP AdvGLUE	Avg GLUE+AdvGLUE	Avg $\Delta\downarrow$
RoBERTa$_{BASE}$											
Vanilla	0%	94.29	24.05	77.91	28.15	92.97	27.43	91.58	19.49	56.98	64.41
R-Drop	0%	95.32	27.84	79.86	31.36	93.30	28.92	91.86	37.44	60.74	58.70
CHILD-TUNING$_D$	70%	94.21	23.82	75.52	16.54	92.36	31.89	91.64	17.95	55.49	65.88
SMART	0%	94.98	35.95	77.54	24.44	93.35	34.29	91.04	46.58	62.27	53.91
FreeLB	0%	94.89	35.81	78.42	32.10	93.12	36.22	92.04	44.10	63.34	52.56
ROSE-First	60%	94.84	37.67	78.34	35.49	92.19	44.19	89.56	44.44	64.59	48.29
ROSE-Second	60%	93.78	36.99	78.16	37.97	92.41	34.63	90.48	45.73	63.77	49.88
ROSE-Ensemble	60%	94.09	39.36	78.63	38.02	92.64	39.59	90.39	47.44	65.02	47.84
RAHP	60%	94.56	38.90	78.95	39.60	92.78	42.19	91.96	48.14	**65.89**	**47.36**
RoBERTa$_{LARGE}$											
Vanilla	0%	96.08	56.08	85.92	61.73	94.58	63.38	92.09	40.60	73.81	36.72
R-Drop	0%	96.59	53.38	85.56	66.67	95.01	55.95	92.35	44.80	73.79	37.18
CHILD-TUNING$_D$	70%	95.91	51.35	85.92	61.73	94.30	58.11	92.03	43.59	72.87	38.35
SMART	0%	96.67	59.12	85.02	69.14	94.91	61.04	92.12	50.85	76.11	32.14
FreeLB	0%	96.49	59.32	86.76	66.91	94.99	62.30	92.60	48.21	75.95	33.53
ROSE-First	60%	95.58	57.77	85.13	70.62	94.08	64.02	90.67	60.26	77.27	28.20
ROSE-Second	60%	96.29	60.59	85.08	67.49	94.72	63.68	91.68	55.90	76.93	30.03
ROSE-Ensemble	60%	96.10	60.81	85.92	71.11	94.26	64.64	91.46	60.51	78.10	27.67
RAHP	60%	96.32	62.02	85.96	71.25	94.74	67.51	90.31	57.12	**78.15**	**27.36**

Table 1: Accuracy on GLUE and AdvGLUE benchmarks, averaged over 5 random seeds. The last column shows the drop from GLUE to AdvGLUE (lower is better). Bold indicates the best result; all baseline results are based on Jiang et al. (2022).

4.3 MAIN RESULTS

Table 1 presents the performance of RAHP compared to a range of competitive baselines across four representative GLUE tasks and their adversarial counterparts in AdvGLUE. Results are reported for both $\text{RoBERTa}_{\text{BASE}}$ and $\text{RoBERTa}_{\text{LARGE}}$, averaged over five random seeds.

For $\text{RoBERTa}_{\text{BASE}}$, RAHP achieves the highest overall average score of 65.89 across GLUE and AdvGLUE, while also attaining the lowest degradation from clean to adversarial performance ($\Delta = 47.36$). This represents a consistent improvement over the ROSE variants, which were the strongest pruning-based baselines. Notably, RAHP narrows the robustness gap without sacrificing performance on the clean GLUE tasks, highlighting its ability to better preserve essential attention heads under heavy pruning (60%).

For $\text{RoBERTa}_{\text{LARGE}}$, the trend is even more pronounced. RAHP reaches a new state of the art with an average score of 78.15, improving upon the best ROSE variant (78.10) while yielding the smallest drop between GLUE and AdvGLUE ($\Delta = 27.36$). This reduction in performance degradation indicates that RAHP not only maintains clean-task accuracy but also enhances resilience to human-crafted adversarial examples.

Figure 2 illustrates the layer-wise distribution of pruned heads in $\text{RoBERTa}_{\text{BASE}}$ on the SST-2 task across different pruning ratios. At a low pruning ratio of 10%, pruning is almost exclusively concentrated in the later layers, particularly from layer 9 onward. Increasing the ratio to 20% preserves this pattern while extending pruning upward to layer 8, suggesting that robustness-aware pruning first targets redundancy in the deepest layers before affecting earlier components.

This trend becomes more pronounced as the pruning ratio grows. For ratios of 40% and above, the distribution follows a clear structural pattern: layers 9–12 consistently absorb the highest pruning rates, often exceeding 80% of their heads. From layer 8 downward, the proportion of pruned heads gradually decreases, reaching a minimum around layers 5–6. Interestingly, pruning volumes then rise again in the earliest layers (layers 1–4), indicating that shallow layers also contain removable redundancy once the deeper layers have been heavily pruned.

Overall, these results reveal a non-uniform pruning distribution: robustness-aware pruning strongly favors removing heads in the deepest layers, followed by shallow layers at high pruning volumes, while the middle layers are relatively more protected. This layered trend aligns with prior observations that attention heads in later layers are more redundant, whereas middle layers encode features more central to task accuracy, as suggested in other papers (Ling et al., 2024; Zhang et al., 2024; Sajjad et al., 2023; Gromov et al., 2024).

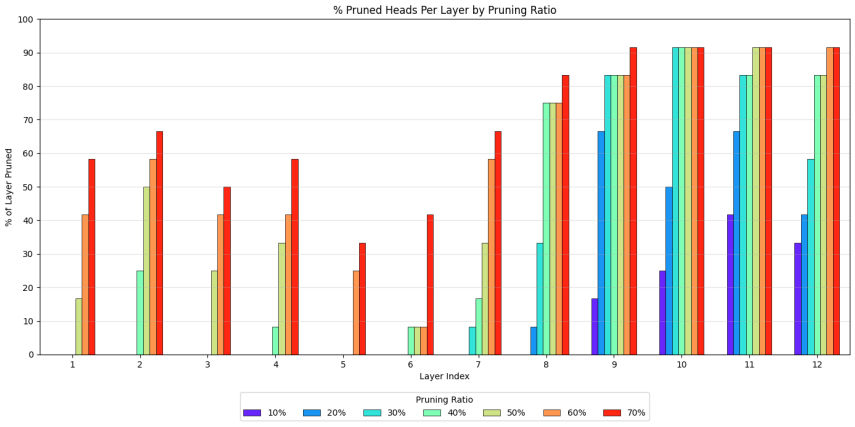

Figure 2: **Distribution of pruned heads per layer in $\text{RoBERTa}_{\text{BASE}}$ (SST-2)** across pruning ratios.

4.4 Analysis of Pruning Behavior

To better understand RAHP's internal behavior, we visualize the per–head Fisher and CLEVER scores and examine how they interact in pruning. All results below are reported for RoBERTa$_{\text{BASE}}$.

Figure 3 shows side-by-side heatmaps of the normalized Fisher importance values and normalized ΔCLEVER scores. Each cell corresponds to an attention head, indexed by its layer (rows) and head position (columns). The Fisher heatmap highlights heads with the largest gradient-based sensitivity, while the ΔCLEVER heatmap indicates robustness changes when individual heads are removed (brighter values correspond to larger robustness gains). Interestingly, Fisher values reveal clear structural patterns: for instance, head 10 consistently exhibits high Fisher importance across layers, suggesting it encodes strongly loss-sensitive features shared throughout the network. By contrast, the ΔCLEVER map highlights different subsets of heads whose removal improves robustness, particularly concentrated in mid-to-late layers. This contrast underscores the complementary nature of the two metrics: Fisher emphasizes heads that are critical for loss optimization and therefore important for preserving model accuracy, whereas ΔCLEVER identifies heads most relevant for robustness gain. In practice, this complementarity ensures that RAHP does not overfit to a single criterion but instead balances accuracy preservation with robustness gains. Thus, we aim to retain heads that appear *dark* in the Fisher heatmap, as these indicate high importance for preserving accuracy, while removing heads that appear *light* in the ΔCLEVER heatmap, as these mark positions whose removal contributes to robustness gains.

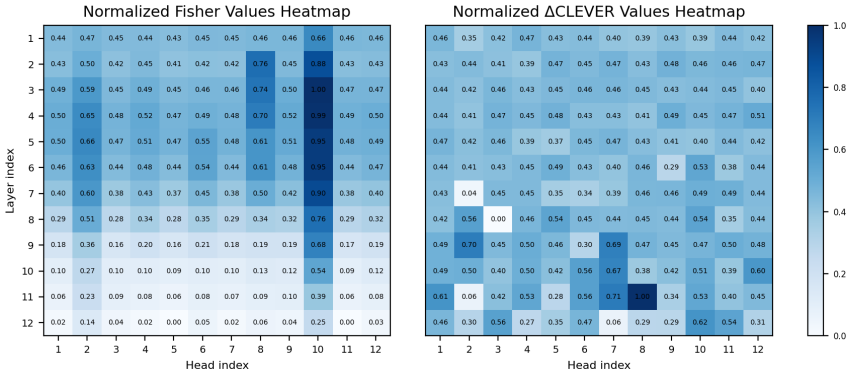

Figure 3: **Normalized Fisher and ΔCLEVER heatmaps for RoBERTa$_{\text{BASE}}$, on SST-2.** Fisher highlights loss-sensitive heads, while ΔCLEVER emphasizes robustness-critical heads in mid-to-late layers, showing their complementary roles in RAHP.

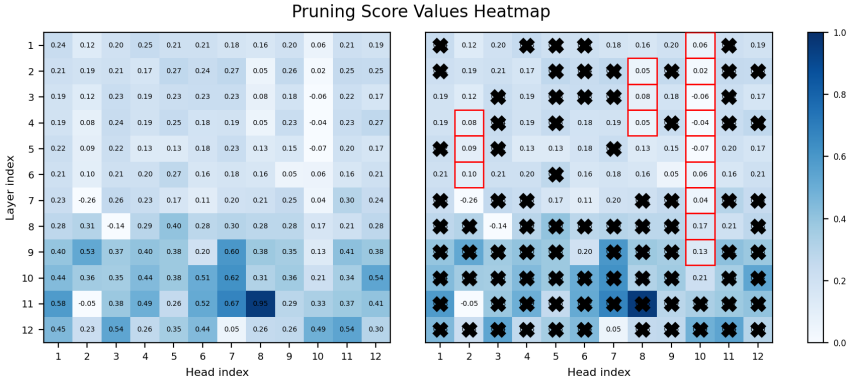

Figure 4: **Composite pruning scores and RAHP decisions for RoBERTa$_{\text{BASE}}$, on SST-2.** Left: composite score matrix ($\beta = 1$, $\alpha = 0.5$). Right: pruning outcomes, where black crosses mark pruned heads and red boxes denote the 10% non-prunable set.

Figure 4 illustrates the composite pruning scores obtained after combining $\beta = 1$ with $\alpha = 0.5$. The left panel shows the full composite score matrix, while the right panel overlays RAHP's pruning decisions: black crosses mark heads selected for pruning, and red boxes denote heads protected as part of the 10% non-prunable fraction. A clear pattern emerges: most pruning decisions occur in the deeper layers, where many heads exhibit higher composite scores. This trend aligns with prior findings that middle-to-deeper attention heads are more redundant and can be pruned with minimal performance loss (Ling et al., 2024; Zhang et al., 2024; Sajjad et al., 2023; Gromov et al., 2024).

5 Ablation Study

To assess the impact of our scoring function's trade-off parameters, we conduct an ablation study varying the values of α (Fisher-based accuracy cost) and β (robustness reward via ΔCLEVER). While our earlier experiments surveyed two other methods, here we focus on two representative models: DeBERTaV3 and DistilBERT, to illustrate how both a high-capacity variant and a lightweight distilled model respond to different (α, β) settings. Figure 5 visualizes the results for two models: ***DeBERTaV3*** and ***DistilBERT***. Each point represents a specific (α, β) configuration, evaluated by its clean accuracy and CLEVER robustness score.

The results reveal that neither a purely accuracy-driven objective nor a purely robustness-driven objective yields a desirable trade-off. For instance, while α=0, β=1 achieves the highest CLEVER score, it suffers from a significant drop in accuracy. Conversely, α=1, β=0 maintains high accuracy but offers minimal robustness gains. The configuration α=0.5, β=1 consistently strikes an effective balance across both models, and among all other models presented during the paper, significantly improving robustness over the baseline while preserving accuracy. We adopt this setting for all main experiments, confirming the need to balance accuracy and robustness during pruning.

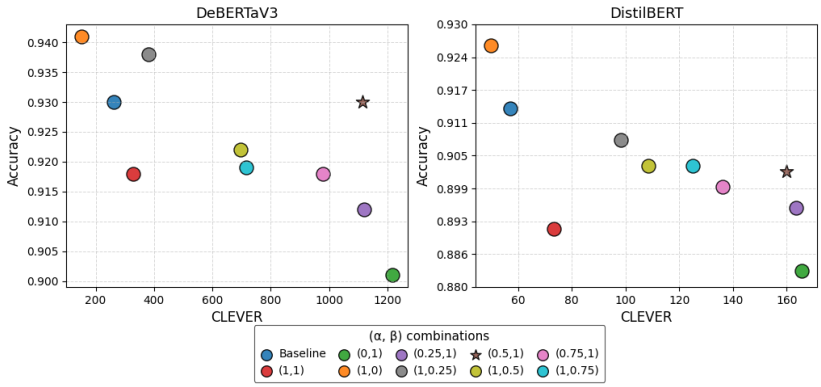

Figure 5: **Effect of** (α, β) **weights** on the robustness–accuracy trade-off for DeBERTaV3 (left) and DistilBERT (right). The star marks our recommended $(0.5, 1)$ setting.

6 Conclusions

We proposed RAHP, a robustness-aware head pruning framework that balances certified robustness and efficiency in Transformers. By combining Fisher-based accuracy costs with CLEVER-based robustness rewards, RAHP prunes heads that minimally affect accuracy while improving robustness, achieving a principled trade-off absent in prior pruning approaches. Experiments on GLUE and AdvGLUE show RAHP consistently surpasses strong baselines, narrowing the robustness gap while maintaining clean-task accuracy. Analysis further reveals pruning concentrates in deeper and, at higher volumes, also shallow layers, consistent with recent findings on Transformer redundancy.

These results suggest that robustness-aware structural pruning offers a practical and scalable path toward certifiably robust Transformers, reducing model size without costly adversarial retraining. Future work may extend RAHP to other modalities (e.g., vision or speech) and explore adaptive pruning ratios.

REFERENCES

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. The fifth pascal recognizing textual entailment challenge. *TAC*, 7(8):1, 2009.

Davis Brown, Charles Godfrey, Cody Nizinski, Jonathan Tu, and Henry Kvinge. Robustness of edited neural networks. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023.

Tianlong Chen, Huan Zhang, Zhenyu Zhang, Shiyu Chang, Sijia Liu, Pin-Yu Chen, and Zhangyang Wang. Linearity grafting: Relaxed neuron pruning helps certifiable robustness. In *International conference on machine learning*, pp. 3760–3772. PMLR, 2022.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pp. 1310–1320. PMLR, 2019.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. Hotflip: White-box adversarial examples for text classification. *arXiv preprint arXiv:1712.06751*, 2017.

Timo Freiesleben and Thomas Grote. Beyond generalization: a theory of robustness in machine learning. *Synthese*, 202(4):109, 2023.

Yujian Gan, Xinyun Chen, Qiuping Huang, Matthew Purver, John R Woodward, Jinxia Xie, and Pengsheng Huang. Towards robustness of text-to-sql models against synonym substitution. *arXiv preprint arXiv:2106.01065*, 2021.

Andrey Gromov, Kushal Tirumala, Hassan Shapourian, Paolo Glorioso, and Daniel A Roberts. The unreasonable ineffectiveness of the deeper layers, 2024. *URL https://arxiv. org/abs/2403.17887*, 2024.

Po-Sen Huang, Robert Stanforth, Johannes Welbl, Chris Dyer, Dani Yogatama, Sven Gowal, Krishnamurthy Dvijotham, and Pushmeet Kohli. Achieving verified robustness to symbol substitutions via interval bound propagation. *arXiv preprint arXiv:1909.01492*, 2019.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. *arXiv preprint arXiv:1911.03437*, 2019.

Lan Jiang, Hao Zhou, Yankai Lin, Peng Li, Jie Zhou, and Rui Jiang. Rose: Robust selective fine-tuning for pre-trained language models. *arXiv preprint arXiv:2210.09658*, 2022.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *Advances in Neural Information Processing Systems*, 36:27469–27500, 2023.

Aounon Kumar, Alex Levine, Tom Goldstein, and Soheil Feizi. Provable robustness against wasserstein distribution shifts via input randomization. *ICLR 2023*, 2023.

Woosuk Kwon, Sehoon Kim, Michael W Mahoney, Joseph Hassoun, Kurt Keutzer, and Amir Gholami. A fast post-training pruning framework for transformers. *Advances in Neural Information Processing Systems*, 35:24101–24116, 2022.

Gui Ling, Ziyang Wang, and Qingwen Liu. Slimgpt: Layer-wise structured pruning for large language models. *Advances in Neural Information Processing Systems*, 37:107112–107137, 2024.

Liyang Liu, Shilong Zhang, Zhanghui Kuang, Aojun Zhou, Jing-Hao Xue, Xinjiang Wang, Yimin Chen, Wenming Yang, Qingmin Liao, and Wayne Zhang. Group fisher pruning for practical network compression. In *International Conference on Machine Learning*, pp. 7021–7032. PMLR, 2021.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. Pruning convolutional neural networks for resource efficient inference. *arXiv preprint arXiv:1611.06440*, 2016.

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

Quora. Quora question pairs dataset. Kaggle, 2018. Retrieved from https://www.kaggle.com/c/quora-question-pairs.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

Andrew Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Hassan Sajjad, Fahim Dalvi, Nadir Durrani, and Preslav Nakov. On the effect of dropping layers of pre-trained transformer models. *Computer Speech & Language*, 77:101429, 2023.

Dvir Samuel and Gal Chechik. Distributional robustness loss for long-tail learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9495–9504, 2021.

Vikash Sehwag, Shiqi Wang, Prateek Mittal, and Suman Jana. Hydra: Pruning adversarially robust neural networks. *Advances in Neural Information Processing Systems*, 33:19655–19666, 2020.

Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of vision transformers. *arXiv preprint arXiv:2103.15670*, 2021.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.

Yi-Lin Sung, Varun Nair, and Colin A Raffel. Training neural networks with fixed sparse masks. *Advances in Neural Information Processing Systems*, 34:24193–24205, 2021.

Lucas Theis, Iryna Korshunova, Alykhan Tejani, and Ferenc Huszár. Faster gaze prediction with dense networks and fisher pruning. *arXiv preprint arXiv:1801.05787*, 2018.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. *arXiv preprint arXiv:2111.02840*, 2021.

Tsui-Wei Weng, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho-Jui Hsieh, and Luca Daniel. Evaluating the robustness of neural networks: An extreme value theory approach. *arXiv preprint arXiv:1801.10578*, 2018.

Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International conference on machine learning*, pp. 5286–5295. PMLR, 2018.

Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. R-drop: Regularized dropout for neural networks. *Advances in neural information processing systems*, 34:10890–10905, 2021.

Runxin Xu, Fuli Luo, Zhiyuan Zhang, Chuanqi Tan, Baobao Chang, Songfang Huang, and Fei Huang. Raise a child in large language model: Towards effective and generalizable fine-tuning. *arXiv preprint arXiv:2109.05687*, 2021.

Yi Zeng, Zhouxing Shi, Ming Jin, Feiyang Kang, Lingjuan Lyu, Cho-Jui Hsieh, and Ruoxi Jia. Towards robustness certification against universal perturbations. In *International Conference on Learning Representation*. ICLR, 2023.

Yang Zhang, Yawei Li, Xinpeng Wang, Qianli Shen, Barbara Plank, Bernd Bischl, Mina Rezaei, and Kenji Kawaguchi. Finercut: Finer-grained interpretable layer pruning for large language models. *arXiv preprint arXiv:2405.18218*, 2024.

LI Zhangheng, Tianlong Chen, Linyi Li, Bo Li, and Zhangyang Wang. Can pruning improve certified robustness of neural networks? *Transactions on Machine Learning Research*, 2022.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. Freelb: Enhanced adversarial training for natural language understanding. *arXiv preprint arXiv:1909.11764*, 2019.