# ICML 2026 Submission #22270
# Entropy Meets Importance: A Unified Head Importance–Entropy Score for Stable and Efficient Transformer Pruning

Table R.1: **Additional Baseline Evaluation** (in response to *Reviewer egVn*, *Reviewer EJiM*). LLaMA-2-7B reasoning benchmark comparison at iso-FLOP budget (matched to HIES 30% head pruning, ≈12% total FLOPs reduction). Reasoning Avg. is the macro-average of HellaSwag, Winogrande, ARC-e, and ARC-c (0-shot).

| **Method** | **Granularity** | **Pruning** | **Weight Restoration** | **Reasoning Avg.** |
|---|---|---|---|---|
| ShortGPT | layer | 4 layers (12.5%) | No Required | 0.493 |
| SparseGPT | weight | 12.95% sparse | Required | 0.544 |
| HIES (ours) | head | 10/32 heads (31.25%) | No Required | 0.534 |

Table R.2: **Systematic Analysis of Preserved Heads** (in response to *Reviewer DBbk*). ERASER faithfulness metrics (comprehensiveness ↑, sufficiency ↓; DeYoung et al., 2020) of heads retained by each pruning method on SST-2 and MNLI. HIES uses a single fixed $\alpha = 0.5$ (no per-task tuning). At 50% pruning, HIES retains heads with strictly better faithfulness than HIS.

| | **HIS** | | **HIES (ours, $\alpha$=0.5)** | |
|---|---|---|---|---|
| **Pruning** | Comp↑ | Suff↓ | Comp↑ | Suff↓ |
| *SST-2* | | | | |
| 0% | 0.2405 | 0.2783 | 0.2405 | 0.2783 |
| 10% | 0.2331 | 0.2804 | 0.2351 | 0.2790 |
| 30% | 0.2328 | 0.2806 | 0.2371 | 0.2797 |
| 50% | 0.1805 | 0.3129 | **0.1907** | **0.3027** |
| *MNLI* | | | | |
| 0% | 0.1018 | 0.2124 | 0.1018 | 0.2124 |
| 10% | 0.1021 | 0.2137 | 0.0994 | 0.2211 |
| 30% | 0.1010 | 0.2267 | 0.0996 | 0.2206 |
| 50% | 0.0955 | 0.2510 | **0.0983** | **0.2352** |

Table R.3: **Ablation on Fusion Strategies** (in response to *Reviewer DBbk*, *Reviewer MGYd*). Accuracy comparison across fusion methods on BERT-base GLUE.

| Ratio | Method | SST-2 | CoLA | MRPC | QQP | STS-B | QNLI | MNLI | RTE | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|
| 10% | Multiplicative fusion | 88.65 | **76.69** | 90.65 | 90.57 | 93.99 | 90.52 | 82.75 | **72.20** | 85.75 |
| | Rank-based fusion | 88.88 | **76.69** | 89.88 | 90.82 | 93.69 | 90.76 | 83.50 | 70.76 | 85.62 |
| | HIES (ours) | **92.66** | 75.48 | **91.04** | **91.93** | **94.00** | **91.03** | **84.04** | 71.84 | **86.50** |
| 30% | Multiplicative fusion | 84.17 | **75.23** | 87.86 | 87.62 | 92.68 | 85.52 | 78.75 | 64.26 | 82.01 |
| | Rank-based fusion | 85.78 | 69.98 | 87.52 | 89.08 | 92.88 | 84.99 | 77.00 | 62.45 | 81.21 |
| | HIES (ours) | **91.86** | 74.97 | **88.81** | **90.37** | **93.89** | **89.13** | **82.50** | **70.04** | **85.20** |
| 50% | Multiplicative fusion | 81.54 | **74.85** | 84.07 | **81.78** | 91.65 | 69.85 | 66.25 | 54.15 | 75.52 |
| | Rank-based fusion | 82.00 | 50.00 | 84.48 | 80.53 | 82.13 | 62.64 | 49.25 | 59.21 | 68.78 |
| | HIES (ours) | **90.71** | 68.52 | **86.80** | 85.73 | **92.65** | **82.68** | **79.00** | **65.34** | **81.43** |

Table R.4: **Ablation on Fusion Strategies** (in response to *Reviewer DBbk*, *Reviewer MGYd*). Stability comparison across fusion methods on BERT-base GLUE.

| Ratio | Method | SST-2 | CoLA | MRPC | QQP | STS-B | QNLI | MNLI | RTE | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|
| 10% | Multiplicative fusion | 94.84 | 97.12 | **94.36** | 96.78 | 82.20 | 95.97 | 95.00 | **94.22** | 93.81 |
| | Rank-based fusion | 96.67 | **97.22** | 93.87 | 96.98 | 79.07 | 95.99 | 94.75 | 93.14 | 93.46 |
| | HIES (ours) | **98.17** | 96.36 | 94.12 | **97.80** | **96.67** | **97.25** | **95.74** | 93.50 | **96.20** |
| 30% | Multiplicative fusion | 85.55 | 90.80 | 86.03 | 88.04 | 66.07 | 83.27 | 81.25 | 85.56 | 83.32 |
| | Rank-based fusion | 89.22 | 82.84 | 85.05 | 89.78 | 61.60 | 87.55 | 88.25 | 83.39 | 83.46 |
| | HIES (ours) | **97.13** | **93.38** | **89.46** | **94.97** | **86.67** | **93.23** | **91.56** | **88.81** | **91.90** |
| 50% | Multiplicative fusion | 82.22 | **88.78** | 76.72 | 84.66 | 23.60 | 69.06 | 76.00 | 67.51 | 71.07 |
| | Rank-based fusion | 82.91 | 81.21 | 79.17 | 82.96 | 27.87 | 63.52 | **79.00** | 66.79 | 70.43 |
| | HIES (ours) | **95.07** | 83.13 | **84.56** | **88.27** | **75.20** | **85.54** | 78.29 | **81.23** | **83.91** |

Table R.5: **Overhead Analysis of Entropy Computation** (in response to *Reviewer DBbk*). Computational overhead of HIES compared to HIS on an NVIDIA RTX 3090, showing that the additional cost is negligible.

| Stage | BERT-base | LLaMA-2-7B |
|---|---|---|
| HIS computation (shared) | $1.49 \pm 0.22$ s | $6.42 \pm 0.07$ s |
| AE computation (HIES additional) | 26.9 ms | 3.19 s |
| $\alpha$ grid search (HIES-only) | 9.37 ms | 65.6 ms |
| **Total HIES additional cost** | **36.3 ms** | **3.26 s** |

Table R.6: **Extended Orthogonality Analysis** (in response to *Reviewer MGYd*, *Reviewer egVN*). Orthogonality of projected gradient directions ($\tilde{u}_h$, $\tilde{v}_h$) across model scales. All values confirm near-orthogonality ($\cos(\theta) = 0.06$–$0.11$, corresponding to $84°$–$87°$), supporting Lemma 4.3.

| Model | # Heads | Mean $\cos(\theta)$ | Std $\cos(\theta)$ |
|---|---|---|---|
| BERT-base | 144 | 0.1102 | 0.0447 |
| LLaMA-2-7B | 1024 | 0.1090 | 0.1365 |
| LLaMA-2-13B | 1600 | 0.0994 | 0.1020 |
| LLaMA-2-70B | 5120 | 0.0574 | 0.0712 |

Table R.7: **Bridging Theoretical Analysis and the HIES Formulation** (in response to *Reviewer MGYd*). Theoretical grounding of HIES, showing how each component follows from the corresponding analytical results.

| **Theoretical result** | **Section** | **Consequence for HIES** |
|---|---|---|
| Loss-increase bound via HIS | §4.2.1, Lemma 1 | Motivates $\widehat{\text{HIS}}_h$ |
| Generalization gap bound via AE | §4.2.2 | Motivates $(1 - \widehat{\text{AE}}_h)$ |
| Composite risk minimization | §4.2.3 | Proves linear HIES is globally optimal |
| Orthogonality of signals | §4.2.4, Lemma 4.3 | Guarantees non-redundancy |

Table R. 8: **Real Hardware Speedup** (in response to *Reviewer egVn*).Inference latency and speedup under head pruning on LLaMA-2-7B, demonstrating larger gains at longer sequence lengths.

| **Model** | **SeqLen** | **Prune%** | **Heads** | **Latency (ms)** | **Speedup** |
|---|---|---|---|---|---|
| LLaMA-2-7B (fp16) | 1024 | 0% | 1024 | $293.57 \pm 1.25$ | 1.00× |
| | | 30% | 704 | $256.55 \pm 1.26$ | 1.14× |
| | | 50% | 512 | $232.33 \pm 0.92$ | 1.26× |
| | 2048 | 0% | 1024 | $683.36 \pm 1.67$ | 1.00× |
| | | 30% | 704 | $570.09 \pm 1.01$ | 1.20× |
| | | 50% | 512 | $505.20 \pm 0.85$ | 1.35× |
| | 4096 | 0% | 1024 | $1586.07 \pm 1.70$ | 1.00× |
| | | 30% | 704 | $1287.03 \pm 1.35$ | 1.23× |
| | | 50% | 512 | $1118.82 \pm 0.62$ | 1.42× |

Table R.9: **GQA Model Evaluation** (in response to *Reviewer egVn*). Reasoning benchmarks on recent GQA-based architectures at 10/30/50% head pruning. A single fixed $\alpha$ 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 R.10: **Recent Model Evaluation.** Reasoning benchmarks on additional recent architectures at 10/30/50% head pruning. A single fixed $\alpha$ is used per model without per-task tuning. HIES outperforms HIS across all pruning ratios on all four models.

| Pruning | Method | HellaSwag | ARC-C | ARC-E | Winogrande | Avg. |
|---|---|---|---|---|---|---|
| Llama-3.1-8B (GQA, 32Q/8KV, $\alpha$=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** |