IDENTIFYING SUPPORT KNOWLEDGE REPRESENTATION FOR LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Attention head pruning provides structured compression for large language models, yet identifying which heads carry target-domain capabilities remains challenging. To initiate the exploration of target-domain knowledge, we define the Support Knowledge (SK), which is the learned patterns of information selection and integration that support target-domain predictions. Based on the definition of SK, we further propose the Activation-Calibrated Margin (ACM) to determine their retention priority. The ACM combines intra-layer rankings of the token competition margin, defined as the difference between the top two attention logits, and the head output magnitude. During scoring, both statistics are collected in a single forward pass over calibration data without training or gradient computation. We then probe expanding candidate sets per layer, allocating a fixed pruning budget across layers based on marginal increases in loss and prediction disagreement proportion for each additional candidate. On domain datasets, ACM better maintains language modeling quality and prediction agreement with the full model than existing baselines, especially under higher pruning ratios, while achieving the highest accuracy in domain question-answering experiments among compared pruning methods.

1 INTRODUCTION

Deploying large language models in specialized domains is constrained by computational and memory budgets. Given the redundancy observed among attention heads (Michel et al., 2019; Voita et al., 2019), head-level pruning provides a natural avenue for structured compression by removing entire computational units. Deployment in specialized domains demands preserving predictions on target distributions, yet head utility varies across domains and tasks (Gong et al., 2021). We therefore address a central question: which attention heads should be retained in a training-free regime using only a compact target-domain calibration corpus?

This problem poses two principal challenges. First, in a training-free setting that permits only forward passes, exhaustive search over head subsets is intractable; moreover, while removing a single head often causes little degradation, removing multiple heads jointly compounds the damage (Michel et al., 2019), invalidating additive cost assumptions. Second, Transformer layers exhibit non-uniform sensitivity to pruning, and uniform per-layer ratios or global ranking can concentrate damage within vulnerable layers (Yin et al., 2024; An et al., 2023).

Mechanistic interpretability has identified heads involved in factual recall, reused across tasks, or carrying task representations (Yao et al., 2024; Merullo et al., 2024; Todd et al., 2024). However, these studies rely on activation patching over paired counterfactual inputs to isolate circuits, yielding no actionable pruning decisions under discrete resource budgets. Conversely, training-free pruning heuristics evaluate removal costs via activation statistics or output fluctuations (An et al., 2023). While computationally efficient, they focus on head outputs, overlooking token selection within the attention mechanism.

Attention token selection can be characterized through max-margin principles (Tarzanagh et al., 2024). Heads with narrow margins between their top-1 and top-2 attention logits operate along competition boundaries, and we hypothesize that such heads are critical for domain-specific predictions. Layer-wise count-matched controls support this hypothesis: removing low-margin heads induces greater predictive degradation on domain text than removing random heads, a disparity that persists when controlling for output magnitude (Section 4.2). Motivated by this insight, we define

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

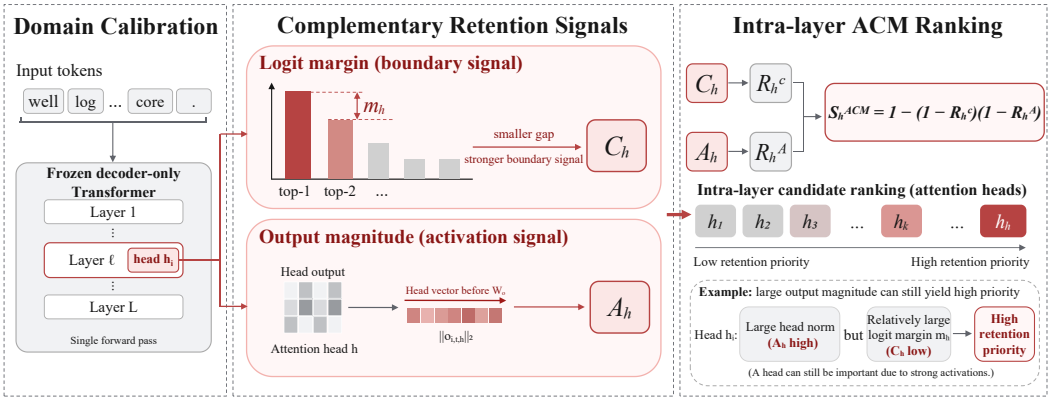

Figure 1: Overview of the ACM scoring pipeline. A single calibration forward pass on the frozen model extracts attention logits and head outputs before the output projection. The token competition margin between the top two logits is exponentially transformed and averaged across positions to form the boundary score, while the average ℓ_2-norm of these head outputs forms the activation score. Converting both scores into intra-layer ranks yields composite ACM scores that establish the candidate pruning order per layer, with inter-layer budgets subsequently allocated via calibration probing.

Support Knowledge (SK) as the learned patterns of information selection and integration that the model repeatedly leverages for domain capability, with attention heads as candidate carriers. We operationalize this definition through Activation-Calibrated Margin (ACM), which captures information selection with token competition margins and information integration with pre-projection head output magnitudes, requiring only a single calibration forward pass (Figure 1). Given these intra-layer rankings, ACM allocates a fixed pruning budget across layers by probing the cumulative predictive degradation of candidate prefixes on calibration data.

Across general and petroleum benchmarks, pruning 30% of attention heads with ACM achieves lower PPL ratios and higher top-1 agreement than competing baselines across all evaluated models and datasets. On downstream domain question-answering, ACM achieves the highest accuracy at 20% and 30% pruning ratios, surpassing FLAP by 1.0 to 7.3 percentage points.

Our contributions are summarized as follows:

- We propose the Activation-Calibrated Margin (ACM) metric to prioritize candidate heads carrying Support Knowledge (SK) for attention head pruning.
- We design a training-free inter-layer budget allocation scheme via marginal predictive damage probing under fixed pruning budgets.

2 Related Work

2.1 Attention Head Importance and Pruning

Pruning attention heads hinges on quantifying each head's contribution to final predictions. Task-driven approaches typically evaluate sensitivity via post-ablation loss shifts, mask gradients, or Fisher information, or alternatively optimize sparse sub-networks through differentiable gating (Michel et al., 2019; Voita et al., 2019; Li et al., 2021; Kwon et al., 2022; Xia et al., 2022; Ma et al., 2023). Although effective, these strategies require backward passes, continuous relaxation, or post-pruning fine-tuning. Conversely, diagnostic methods profile intrinsic attention behaviors, such as peak sharpness, entropy, and concentration, to probe functional specialization and substitutability (Gong et al., 2021; Budhraja et al., 2020; Hassid et al., 2022; Choi et al., 2026). However, these localized statistics do not readily reflect post-pruning predictive changes without accounting for the model's global computation (Budhraja et al., 2020; Wen et al., 2023).

In post-training structured pruning, structural contributions are commonly estimated from activation magnitudes, layer-wise reconstruction errors, or empirical degradation upon component removal (Kurtić et al., 2023; Meng et al., 2024; Cai et al., 2024; Muralidharan et al., 2024; Kolawole et al., 2026). For instance, FLAP assesses post-removal recoverability through feature fluctuations and applies bias compensation to alleviate output drift (An et al., 2023). In contrast, ACM establishes intra-layer head rankings without iterative perturbations or layer-wise reconstruction. During a single forward pass over the frozen model, ACM simultaneously extracts the margin between the top two pre-softmax attention logits and the pre-projection head output magnitude, jointly accounting for token competition intensity and output responsiveness.

2.2 Post-Training Structured Pruning and Budget Allocation

Post-training compression commonly relies on forward activation statistics collected over compact calibration sets. Early formulations, notably Optimal Brain Compression and SparseGPT, compute second-order information to reconstruct layer outputs while updating unpruned weights, whereas Wanda simplifies this selection via the product of weight magnitudes and input activation norms (Frantar & Alistarh, 2022; 2023; Sun et al., 2024). While these techniques predominantly induce unstructured or semi-structured sparsity, structured alternatives compress models along complete computational dimensions. For instance, SliceGPT concentrates channel variance through orthogonal rotations prior to slicing embedding dimensions, and DISP-LLM unbinds hidden-dimension dependencies between adjacent blocks (Ashkboos et al., 2024; Gao et al., 2024). Parallel efforts investigate joint mask learning in Sheared LLaMA, layer-wise reconstruction in SlimGPT, pruning-aware fine-tuning in PAT, and heterogeneous multi-granularity scoring (Xia et al., 2024; Ling et al., 2024; Liu et al., 2025b;a).

When directly pruning pre-trained backbones, component selection and budget assignment often entail substantial computational overhead. ZipLM iteratively eliminates structures by balancing local loss against physical latency, and OSSCAR solves layer-wise combinatorial objectives to identify attention heads and feed-forward neurons (Kurtić et al., 2023; Meng et al., 2024). Similarly, FLAP estimates recoverability from feature fluctuations with bias compensation, whereas Bonsai evaluates module importance through repeated sub-model perturbations (An et al., 2023; Kolawole et al., 2026). For cross-layer budget allocation, existing frameworks rely on heuristic score normalization (FLAP), outlier activation ratios (OWL), runtime-aware progressive pruning (ZipLM), combinatorial search (EvoPress, Týr-the-Pruner), or minimal linear reconstruction error (LSA) (An et al., 2023; Yin et al., 2024; Kurtić et al., 2023; Sieberling et al., 2025; Li et al., 2025; Yang et al., 2026). Unlike these approaches that depend on iterative re-evaluations or layer-isolated heuristics, ACM determines intra-layer candidate rankings from a single forward pass of the intact model by coupling token competition margins with head output magnitudes. It then probes candidate prefixes to distribute discrete layer budgets according to the marginal degradation of prediction fidelity, incorporating a concentration penalty to prevent over-pruning within localized layers.

2.3 Knowledge Circuits and the Functionality of Attention Heads

Prior work characterizes Transformer feed-forward layers as key-value memories and links factual recall to intermediate MLP layers via causal tracing (Geva et al., 2021; Meng et al., 2022). Attention analyses have further identified induction heads that support in-context learning, functional heads in indirect object identification, and circuit components that coordinate with MLPs, are reused across tasks, or transmit task representations (Olsson et al., 2022; Wang et al., 2022; Yao et al., 2024; Merullo et al., 2024; Todd et al., 2024). These studies demonstrate that attention heads participate in specific predictive computations, and ablation-based analyses assess such contributions through output changes. Accordingly, ACM treats heads as candidate pruning units and evaluates pruning outcomes by the degree to which predictions are preserved in the target domain.

2.4 Token Selection in Attention and Implicit Bias

The max-margin perspective originates from studies on the implicit bias of gradient-based optimization. On linearly separable data, gradient descent with exponential-tailed losses such as logistic loss directs parameter trajectories toward max-margin solutions, with the specific bias governed by the model parameterization and optimization geometry (Soudry et al., 2018; Gunasekar et al., 2018).

This perspective has since been extended to attention models: analyses of single-layer prompt attention demonstrate that gradient descent biases softmax attention toward context-relevant tokens (Oymak et al., 2023). Max-margin token selection further formalizes the directional convergence of attention parameters as a margin separation between locally optimal tokens and remaining tokens, expanding to query–key parameterized self-attention (Ataee Tarzanagh et al., 2023; Tarzanagh et al., 2024).

Recent work expands this theory to finite-time sparsification, token retrieval dynamics in next-token prediction, trainable embeddings, and generalized optimization geometries (Vasudeva et al., 2025; Li et al., 2024; Wu et al., 2025; Julistiono et al., 2026). These developments link attention optimization directly to logit separation. ACM extracts the attention logit margin between the top two tokens in fixed pretrained models to capture position-level token competition, validating this signal through count-matched head removals and structured pruning.

3 METHOD

We propose Activation-Calibrated Margin (ACM), which integrates attention logit margins and pre-projection head output magnitudes to establish intra-layer attention head rankings. It then allocates a fixed pruning budget across layers based on the marginal predictive changes induced by candidate prefixes on calibration data.

3.1 ATTENTION LOGIT MARGIN

Under specific theoretical assumptions, max-margin analysis connects token selection in attention to the separation of the selected token from remaining tokens (Ataee Tarzanagh et al., 2023; Tarzanagh et al., 2024). Inspired by this view, we use the attention logit margin to characterize token competition in pretrained models.

We omit the layer index for brevity. Let $\mathcal{C}_t$ denote the set of valid causal prefix positions for query position t, let $a_{i,t,s}^{(h)}$ be the pre-softmax attention logit, and let $\alpha_{i,t,h} = \arg\max_{s \in \mathcal{C}_t} a_{i,t,s}^{(h)}$. Let $\mathcal{D}$ denote the set of query positions in the target-domain calibration set whose causal prefixes contain at least two valid tokens, excluding the initial position of each sequence. For sample i, query position t, and head h, the attention logit margin is defined as the difference between the largest and second-largest logits:

$$m_h(i,t) = \max_{s \in \mathcal{C}_t} a_{i,t,s}^{(h)} - \max_{s \in \mathcal{C}_t \setminus \{\alpha_{i,t,h}\}} a_{i,t,s}^{(h)}, \tag{1}$$

where $(i,t) \in \mathcal{D}$; in case of ties for the maximum logit, the margin is zero. To aggregate low-margin behavior over the calibration set, we map position-level margins to continuous boundary intensities:

$$c_h(i,t) = \exp\left(-\frac{m_h(i,t)}{\tau}\right), \tag{2}$$

where the temperature $\tau > 0$ weights small margins. We set $\tau = 0.5$ by default and evaluate its sensitivity in Section 4.5. The boundary intensity equals one at zero margin and decays exponentially as the margin increases.

3.2 HEAD-LEVEL ACM SCORE

Averaging the position-level boundary intensity across the calibration set $\mathcal{D}$ yields the boundary score for head h:

$$C_h(\mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(i,t) \in \mathcal{D}} c_h(i,t). \tag{3}$$

This average reflects both the frequency of low-margin positions and their proximity to the decision boundary. We use C_h as a retention signal for a head's participation in domain token competition.

The margin characterizes the intensity of token competition, but does not reflect the magnitude of the vector produced by each head. Let $o_{i,t,h}$ denote the output vector of head h at position (i,t),

taken prior to the attention output projection. We compute its average ℓ_2 norm across calibration positions as a complementary retention signal:

$$A_h(\mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(i,t)\in\mathcal{D}} \|o_{i,t,h}\|_2. \tag{4}$$

Because boundary and activation scores differ in scale, we convert each into a normalized intra-layer rank, deferring inter-layer budget allocation to calibration probing. Let H_ℓ be the number of prunable heads in layer ℓ, and let $\rho_\ell^Z(h)$ be the 1-based rank of head h for signal $Z \in \{C, A\}$ sorted in ascending order, with ties broken by head index. The normalized rank is defined as

$$R_h^Z = \frac{\rho_\ell^Z(h) - 1}{\max(H_\ell - 1, 1)}, \qquad Z \in \{C, A\}. \tag{5}$$

Both rankings span $[0, 1]$, reflecting the relative order of their corresponding signals within the layer.

ACM combines the two rankings into a retention score for each head:

$$S_h^{\mathrm{ACM}} = R_h^C + \left(1 - R_h^C\right) R_h^A = 1 - \left(1 - R_h^C\right)\left(1 - R_h^A\right). \tag{6}$$

The activation signal elevates retention scores for heads with low boundary rankings but relatively high output magnitudes within their layer. A low composite score requires both signals to rank low in the layer. We sort heads in ascending order of their ACM scores to establish the candidate pruning order per layer. The individual effects are analyzed in Section 4.5.

3.3 Pruning Application Under a Fixed Budget

While the ACM score establishes intra-layer candidate order, the inter-layer budget is allocated via calibration probing. Restricting candidate pruning sets to prefixes of heads sorted by ascending ACM scores reduces candidate selection to allocating integer budgets across layers. For a model with $H = \sum_\ell H_\ell$ prunable heads and pruning ratio $r \in [0, 1]$, budget allocation determines the candidate prefix length per layer subject to pruning exactly $K = \mathrm{round}(rH)$ heads in total.

Prefix damage estimation. Let $\pi_\ell(j)$ denote the j-th head in layer ℓ sorted by ascending ACM score, breaking ties by head index. For $j = 1, \ldots, H_\ell$, we zero out the output vectors of the first j candidates in layer ℓ while keeping other layers unpruned, recording the per-token NLL increase $N_\ell(j)$ and the top-1 prediction disagreement rate $Q_\ell(j)$ relative to the unpruned model. Setting $N_\ell(0) = Q_\ell(0) = 0$, the adjacent-prefix marginal changes are defined as follows:

$$\begin{aligned} \Delta_{\ell,j}^{\mathrm{NLL}} &= N_\ell(j) - N_\ell(j-1), \\ \Delta_{\ell,j}^{\mathrm{pred}} &= Q_\ell(j) - Q_\ell(j-1). \end{aligned} \tag{7}$$

All candidate prefixes share the same calibration probing subset.

Marginal cost and budget allocation. Applying unit-weight isotonic regression to both marginal sequences within each layer and normalizing fitted values across layers, the marginal cost of the j-th candidate in layer ℓ is

$$\begin{aligned} d_{\ell,j} = &\frac{\widehat{\Delta}_{\ell,j}^{\mathrm{NLL}}}{s_{\mathrm{NLL}}} + \frac{\widehat{\Delta}_{\ell,j}^{\mathrm{pred}}}{s_{\mathrm{pred}}} \\ &+ \lambda \frac{j-1}{\max(H_\ell - 1, 1)}. \end{aligned} \tag{8}$$

where $\widehat{\Delta}$ denotes fitted marginal changes, $s_{\mathrm{NLL}}, s_{\mathrm{pred}} > 0$ are scale factors, and $\lambda \geq 0$ penalizes concentrated pruning. We select the K lowest costs across layers; Appendix A details the settings and optimality argument for the surrogate.

4 Experiments

4.1 Experimental Setup

Models and Datasets In this section, we focus our analysis on representative models spanning diverse parameter scales and architectural families: Qwen3-14B, Qwen3-8B, Qwen3-1.7B (Yang et al., 2025), Qwen3.8-27B (Qwen Team, 2026), and Llama-3.2-3B (Meta, 2024).

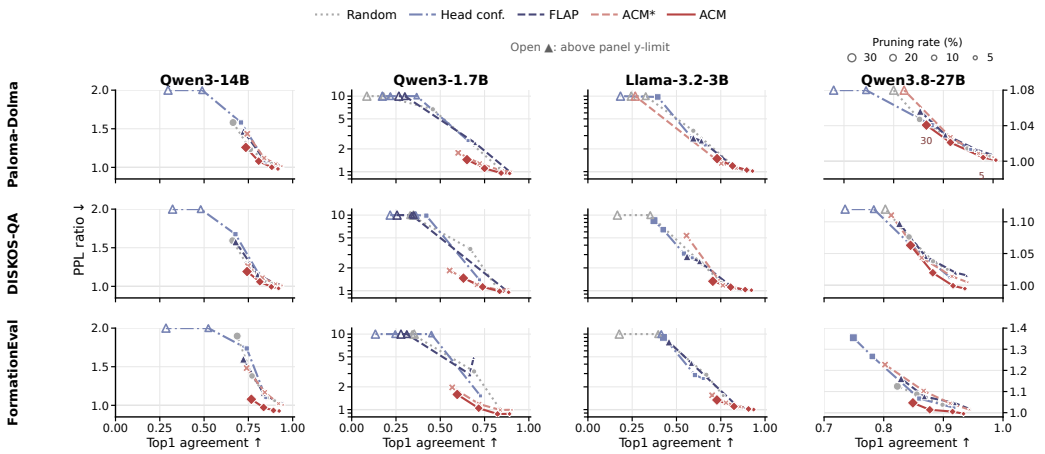

Figure 2: Language modeling quality and predictive fidelity across four models (columns) and three datasets (rows): Paloma-Dolma, DISKOS-QA, and FormationEval. Operating points closer to the bottom-right indicate better retention on both metrics (top-1 agreement vs. PPL ratio). Node sizes scale with pruning ratio (5%, 10%, 20%, 30%), and lines connect each method across ratios. Points exceeding the upper boundary are plotted as open triangles.

For our main evaluation on language modeling and prediction fidelity, we use Paloma-Dolma (Magnusson et al., 2023) as a general-domain corpus alongside DISKOS-QA (georgeghon & contributors, 2026; FORCE NETWORK Group, 2024) and FormationEval (Ermilov, 2026) for specialized vertical domains. To evaluate domain task capabilities, we test on the PetroBench English SCQ balanced subset and FormationEval full MCQ. For PetroBench, English texts supply calibration data for head ranking, while test questions measure multiple-choice performance retention. Both benchmarks score candidate labels via length-normalized NLL, selecting the option with the lowest NLL as the prediction. Complete evaluation results across all models and datasets appear in Appendix C.

Baselines We compare ACM against the following baselines: Random, which prunes heads uniformly at random; HeadConf, which uses the top-1 confidence of attention probabilities as head importance; and FLAP, which constructs pruning scores using head output variations and output projection weights.

Pruning Configurations Our full method incorporates the inter-layer budget allocation described in Section 3.3. To decouple head scoring from budget allocation, we introduce an ablation variant, ACM*, which retains ACM scores but globally prunes the K lowest-scoring heads. Comparing ACM* with other baselines evaluates scoring efficacy under identical global rules, while comparing the full method with ACM* isolates the gain from inter-layer allocation. In both variants, higher scores reflect higher retention priority as candidate carrier components of SK.

Metrics We evaluate pruning via two complementary metrics: the PPL ratio, defined as the pruned model's perplexity normalized by that of the unpruned baseline (lower values indicate better distribution preservation), and top-1 agreement, measuring the proportion of identical next-token predictions between pruned and unpruned models (higher values reflect greater decision fidelity). Together, they capture both global distribution quality and token-level prediction retention.

4.2 Margin Disparities and Predictive Contributions

In a layer-wise count-matched setup across three models and three domain datasets, pruning low-margin heads yields higher PPL ratios in 101/135 cases and lower top1 agreement in 106/135 cases compared to random head removal. Controlling for magnitude confounding via intra-layer stratification, the corresponding proportions are 106/135 and 107/135. Experimental design, complete per-setting breakdowns, and high-margin head analyses appear in Appendix B.

Table 1: Downstream accuracy on the PetroBench English SCQ balanced evaluation subset and FormationEval full MCQ. Full denotes the unpruned model; both benchmarks employ the candidate label length-normalized likelihood scoring protocol.

Model	Pruned	Full	Random	HeadConf	FLAP	ACM
PetroBench English SCQ						
Qwen3-8B	5%	0.776	0.751	0.761	0.764	**0.766**
	10%	0.776	0.732	0.646	0.760	**0.766**
	20%	0.776	0.575	0.258	0.727	**0.737**
	30%	0.776	0.392	0.245	0.631	**0.687**
Qwen3-14B	5%	0.792	0.779	0.751	0.784	**0.798**
	10%	0.792	0.750	0.625	**0.789**	**0.789**
	20%	0.792	0.661	0.275	0.758	**0.780**
	30%	0.792	0.430	0.249	0.641	**0.710**
FormationEval full MCQ						
Qwen3-8B	5%	0.855	0.802	0.780	0.850	**0.859**
	10%	0.855	0.711	0.568	**0.836**	**0.836**
	20%	0.855	0.527	0.288	0.780	**0.813**
	30%	0.855	0.312	0.285	0.692	**0.727**
Qwen3-14B	5%	0.854	0.824	0.760	0.846	**0.859**
	10%	0.854	0.790	0.646	0.838	**0.850**
	20%	0.854	0.663	0.296	0.786	**0.833**
	30%	0.854	0.380	0.250	0.735	**0.808**

4.3 In-Domain Language Modeling and Prediction Fidelity

These comparisons confirm that low-margin heads contribute to domain predictions beyond chance, while pruning high-margin heads likewise causes marked damage (Appendix B). Integrating both signals into a retention score, ACM achieves lower PPL ratios than Random, HeadConf, and FLAP across all 48 configurations spanning four models, three datasets, and four pruning ratios, while surpassing all three external baselines in top-1 agreement in 36 settings. This joint advantage widens at higher pruning ratios: when pruning 30% of the attention heads, ACM achieves a lower PPL ratio and simultaneously achieves higher top-1 agreement across all 12 model-dataset combinations, as shown in Figure 2. Even without inter-layer budget allocation, ACM* surpasses FLAP on both metrics in 30 out of the 48 settings.

On FormationEval with Llama-3.2-3B, pruning 30% of attention heads yields a PPL ratio of 1.34 for ACM versus 7.86 for FLAP, with top1 agreements of 0.728 and 0.458, respectively. While FLAP degrades sharply as pruning intensifies, ACM remains in the high-fidelity region, demonstrating that it simultaneously mitigates degradation in language modeling and prediction consistency.

Comparing ACM with ACM* isolates the benefit of inter-layer budget allocation. On DISKOS-QA, where ACM*'s PPL ratio rises sharply under aggressive pruning, ACM maintains lower PPL ratios across all four models and generally higher agreement, confirming that budget allocation mitigates high-pruning degradation. Full single-metric curves are provided in Appendix C.

To evaluate deployment benefits, we physically prune Llama-3.2-3B by slicing the corresponding rows from query projections and columns from attention output projections dictated by the inter-layer budget allocation. On DISKOS-QA, pruning 20% and 30% of query heads reduces total model parameters by 3.28% and 4.94%, respectively. In input processing benchmarks (batch size 4, sequence lengths 1024 and 4096), these two pruning ratios achieve speedups of $1.023/1.041$ and $1.044/1.073$. These results confirm that full ACM yields modest speedups under long-input regimes (experimental details in Appendix H).

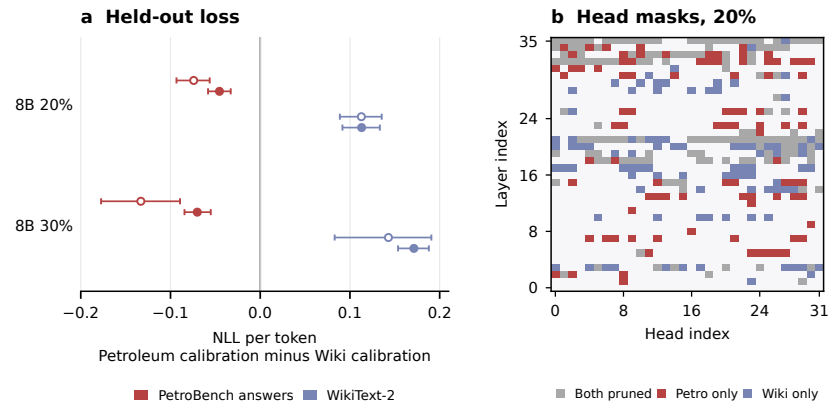

Figure 3: Calibration corpus transfer and head selection on Qwen3-8B. (a) Difference in per-token NLL increase relative to the unpruned baseline between masks calibrated on petroleum text and WikiText-2; negative values indicate that domain-matched (PetroBench) calibration incurs less degradation. (b) Layer- and head-wise pruning masks at a 20% pruning ratio for the first calibration window. Full protocols are given in Appendix F.

4.4 Support Knowledge and Domain Task Preservation

Support Knowledge (SK) entails three testable properties: domain criticality, persistent reliance during inference, and target-domain specificity, which we validate sequentially via downstream question-answering accuracy, candidate consistency across calibration subsets, and cross-corpus transfer. First, we evaluate domain task accuracy to test whether pruning with ACM maintains predictive capabilities associated with SK. On PetroBench and FormationEval, we evaluate Qwen3-8B and Qwen3-14B, both exhibiting high unpruned accuracy, to compare pruning methods under matched pruning ratios against unpruned references (Table 1).

Across all eight configurations spanning two domain tasks and two models, ACM achieves the highest accuracy at both 20% and 30% pruning ratios. ACM outperforms FLAP by 1.0 to 6.9 percentage points on PetroBench and 3.3 to 7.3 percentage points on FormationEval. These advantages widen at 30% pruning across all settings. These demonstrate that ACM preserves domain task performance beyond language modeling, aligning with its design to prioritize candidate carrier components of SK. Full results appear in the Appendix D.

Consistency of Candidate Selection Given that SK reflects computational patterns repeatedly relied upon for domain predictions, we evaluate whether ACM consistently recovers candidate carrier heads across calibration samples. On Qwen3-8B and Qwen3-14B, the mean pairwise Spearman rank correlation of full-model head rankings across three calibration subsets is 0.989 for both models, with mean pairwise Jaccard similarities of 0.871 and 0.901 among the top 10% candidates. Crucially, candidates shared across all three subsets account for 90.4% and 92.5% of each individual set, confirming that ACM stably identifies a recurring core of candidate heads. Pairwise breakdowns are reported in Appendix Table 9.

Calibration Corpus and Attention Head Selection We examine whether the calibration corpus influences ACM's identification of candidate carrier heads and its subsequent prediction retention on held-out text. On Qwen3-8B, across two disjoint calibration windows and pruning ratios of 20% and 30%, we observe consistent domain matching: masks calibrated on PetroBench incur smaller per-token NLL increases on petroleum evaluation text, whereas WikiText-2-derived masks better preserve general test performance, as illustrated in Figure 3(a). At a 20% pruning ratio, the overlap between pruned head sets derived from the two corpora is 60.0% and 55.7% across the two windows, indicating that while calibration corpora steer head selection, the masks share a stable structural core (layer- and head-wise distributions visualized in Figure 3(b)). Because the two corpora differ

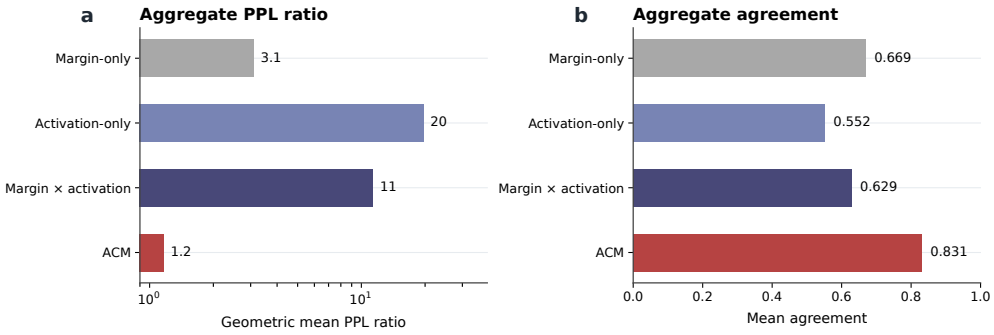

Figure 4: Performance of four scoring variants across 144 paired configurations: margin-only, activation-only, their multiplicative combination, and ACM. Left: geometric mean PPL ratio (lower is better); right: mean top-1 agreement (higher is better).

across topic, vocabulary, and style, this alignment indicates domain matching rather than conclusive evidence of factual knowledge transfer. Experimental protocols are detailed in Appendix F.

4.5 Ablation Studies

To analyze the contributions of the attention logit margin, the head output magnitude, and their combination rule in ACM, we evaluate four scoring formulations across 144 paired configurations. As shown in Figure 4, ACM achieves the strongest overall performance, attaining a geometric mean PPL ratio of 1.17 and a mean top1 agreement of 0.831. See Appendix I for more details.

To evaluate the choice of the default temperature, we vary τ from 0.125 to 2.0 across several models and datasets. Experimental results show that the optimal temperature for both PPL ratio and top1 agreement varies across configurations. With fixed $\tau = 0.5$, the PPL ratio exhibits an average gap of 0.441% and a maximum gap of 1.883% relative to the per-setting optimal values, both representing the lowest deviations among the seven candidate temperatures. For top1 agreement, the average and maximum gaps relative to the optimal values are 0.229 and 1.188 percentage points, respectively. Full results are provided in Appendix J.

5 Conclusion

To preserve target-domain capabilities under structured pruning, we propose Activation-Calibrated Margin (ACM) to evaluate attention head retention priorities. The ACM characterizes attention selection boundaries via token competition margins and output intensity via head output magnitudes, corresponding directly to information selection and integration in our definition of SK. The ACM integrates both signals through intra-layer ranking and allocates inter-layer pruning budgets via marginal predictive damage probing. Comparative experiments demonstrate that across most settings, pruning low-margin heads degrades domain predictions more severely than random removal, a pattern that persists when controlling for output magnitude; concurrently, removing heads with large margins also incurs marked damage in certain settings. Across 48 combinations of models, datasets, and pruning ratios, ACM consistently achieves lower PPL ratios than external baselines, while yielding higher top1 agreement in 36 settings and attaining the highest post-pruning accuracy across 8 domain question-answering settings. Ablations confirm the synergy of both signals, and evaluations across calibration subsets demonstrate ranking stability, corroborating ACM as an effective priority metric for head pruning. Future work could examine whether candidate heads carry specific factual content via targeted interventions, adopt target-domain capability preservation as a unified objective to compare pruning damage across heterogeneous components such as attention heads and MLPs, and explore cross-module budget allocation.

AI USE STATEMENT

In this work, we did not use generative AI tools for any tasks with required disclosure. Additionally, we used generative AI tools to edit the manuscript for grammar, clarity, and readability. We have reviewed all AI-assisted work. The authors checked the AI-edited text to ensure that it preserved the intended meaning and did not alter the scientific claims. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Yongqi An, Xu Zhao, Tao Yu, Ming Tang, and Jinqiao Wang. Fluctuation-based adaptive structured pruning for large language models, 2023. URL https://arxiv.org/abs/2312.11983.

Saleh Ashkboos, Maximilian L. Croci, Marcelo Gennari do Nascimento, Torsten Hoefler, and James Hensman. Slicegpt: Compress large language models by deleting rows and columns, 2024. URL https://arxiv.org/abs/2401.15024.

Davoud Ataee Tarzanagh, Yingcong Li, Xuechen Zhang, and Samet Oymak. Max-margin token selection in attention mechanism. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 48314–48362. Curran Associates, Inc., 2023. doi: 10.52202/075280-2096. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/970f59b22f4c72aec75174aae63c7459-Paper-Conference.pdf.

Aakriti Budhraja, Madhura Pande, Preksha Nema, Pratyush Kumar, and Mitesh M. Khapra. On the weak link between importance and prunability of attention heads. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3230–3235, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.260. URL https://aclanthology.org/2020.emnlp-main.260/.

Ruisi Cai, Saurav Muralidharan, Greg Heinrich, Hongxu Yin, Zhangyang Wang, Jan Kautz, and Pavlo Molchanov. Flextron: Many-in-one flexible large language model. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5298–5311. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/cai24e.html.

Minsik Choi, Hyegang Son, Changhoon Kim, and Young Geun Kim. Entropy meets importance: A unified head importance-entropy score for stable and efficient transformer pruning, 2026. URL https://arxiv.org/abs/2510.13832.

Almaz Ermilov. Formationeval, an open multiple-choice benchmark for petroleum geoscience, 2026. URL https://arxiv.org/abs/2601.02158.

FORCE NETWORK Group. Large oil and gas industry text dataset from norwegian, uk and dutch public oil and gas documents, 2024. URL https://doi.org/10.5281/zenodo.10775273.

Elias Frantar and Dan Alistarh. Optimal brain compression: A framework for accurate post-training quantization and pruning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 4475–4488. Curran Associates, Inc., 2022. doi: 10.52202/068431-0323. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/1caf09c9f4e6b0150b06a07e77f2710c-Paper-Conference.pdf.

Elias Frantar and Dan Alistarh. SparseGPT: Massive language models can be accurately pruned in one-shot. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10323–10337. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/frantar23a.html.

Shangqian Gao, Chi-Heng Lin, Ting Hua, Tang Zheng, Yilin Shen, Hongxia Jin, and Yen-Chang Hsu. Disp-llm: Dimension-independent structural pruning for large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 72219–72244. Curran Associates, Inc., 2024. doi: 10.52202/079017-2305. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/84a7fc24ed52e8eff514c33e8ac76ea3-Paper-Conference.pdf.

georgeghon and contributors. Diskos-qa. https://github.com/georgeghon/DISKOS-QA, 2026.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446/.

Hongyu Gong, Yun Tang, Juan Pino, and Xian Li. Pay better attention to attention: Head selection in multilingual and multi-domain sequence modeling. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 2668–2681. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/15c00b5250ddedaabc203b67f8b034fd-Paper.pdf.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1832–1841. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/gunasekar18a.html.

Michael Hassid, Hao Peng, Daniel Rotem, Jungo Kasai, Ivan Montero, Noah A. Smith, and Roy Schwartz. How much does attention actually attend? questioning the importance of attention in pretrained transformers. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1403–1416, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.101. URL https://aclanthology.org/2022.findings-emnlp.101/.

Addison Kristanto Julistiono, Davoud Ataee Tarzanagh, and Navid Azizan. Optimizing attention with mirror descent: Generalized max-margin token selection. *Journal of Machine Learning Research*, 27(36):1–61, 2026. URL http://jmlr.org/papers/v27/25-0549.html.

Steven Kolawole, Lucio Dery, Jean-François Kagy, Virginia Smith, Graham Neubig, and Ameet Talwalkar. Everybody prune now: Structured pruning of llms with only forward passes, 2026. URL https://arxiv.org/abs/2402.05406.

Eldar Kurtić, Elias Frantar, and Dan Alistarh. Ziplm: Inference-aware structured pruning of language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 65597–65617. Curran Associates, Inc., 2023. doi: 10.52202/075280-2863. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/ced46a50befedcb884ccf0cbe8c3ad23-Paper-Conference.pdf.

Woosuk Kwon, Sehoon Kim, Michael Mahoney, Joseph Hassoun, Kurt Keutzer, and Amir Gholami. A fast post-training pruning framework for transformers. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24101–24116. Curran Associates, Inc., 2022. doi: 10.52202/068431-1750. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/987bed997ab668f91c822a09bce3ea12-Paper-Conference.pdf.

Guanchen Li, Yixing Xu, Zeping Li, Ji Liu, Xuanwu Yin, Dong Li, and Emad Barsoum. Týr-the-pruner: Structural pruning llms via global sparsity distribution optimization. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 149163–149192. Curran Associates, Inc., 2025. doi: 10.52202/085713-4989. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/db2cbf43a349bc866111e791b58c7bf4-Paper-Conference.pdf.

Jiaoda Li, Ryan Cotterell, and Mrinmaya Sachan. Differentiable subset pruning of transformer heads. *Transactions of the Association for Computational Linguistics*, 9:1442–1459, 2021. doi: 10.1162/tacl_a_00436. URL https://aclanthology.org/2021.tacl-1.86/.

Yingcong Li, Yixiao Huang, Muhammed E. Ildiz, Ankit Singh Rawat, and Samet Oymak. Mechanics of next token prediction with self-attention. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 685–693. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/li24f.html.

Gui Ling, Ziyang Wang, Yuliang Yan, and Qingwen Liu. Slimgpt: Layer-wise structured pruning for large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 107112–107137. Curran Associates, Inc., 2024. doi: 10.52202/079017-3401. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/c1c44e46358e0fb94dc94ec495a7fb1a-Paper-Conference.pdf.

Jun Liu, Zhenglun Kong, Pu Zhao, Changdi Yang, Xuan Shen, Hao Tang, Geng Yuan, Wei Niu, Wenbin Zhang, Xue Lin, Dong Huang, and Yanzhi Wang. Toward adaptive large language models structured pruning via hybrid-grained weight importance assessment. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025a. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i18.34078. URL https://doi.org/10.1609/aaai.v39i18.34078.

Yijiang Liu, Huanrui Yang, Youxin Chen, Rongyu Zhang, Miao Wang, Yuan Du, and Li Du. Pat: Pruning-aware tuning for large language models, 2025b. URL https://arxiv.org/abs/2408.14721.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. Llm-pruner: On the structural pruning of large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 21702–21720. Curran Associates, Inc., 2023. doi: 10.52202/075280-0950. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/44956951349095f74492a5471128a7e0-Paper-Conference.pdf.

Ian Magnusson, Akshita Bhagia, Valentin Hofmann, Luca Soldaini, A. Jha, Oyvind Tafjord, Dustin Schwenk, Pete Walsh, Yanai Elazar, Kyle Lo, Dirk Groeneveld, Iz Beltagy, Hanna Hajishirzi, Noah A. Smith, Kyle Richardson, and Jesse Dodge. Paloma: A benchmark for evaluating language model fit. *ArXiv*, abs/2312.10523, 2023. URL https://api.semanticscholar.org/CorpusID:266348815.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Xiang Meng, Shibal Ibrahim, Kayhan Behdin, Hussein Hazimeh, Natalia Ponomareva, and Rahul Mazumder. OSSCAR: One-shot structured pruning in vision and language models with combinatorial optimization. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 35354–35377. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/meng24a.html.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Circuit component reuse across tasks in transformer language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 18349–18377, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/4ffbfbcf4ad7304d57158b046525e46c-Paper-Conference.pdf.

Meta. Llama 3.2 Model Card. Model card, 2024. URL https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD.md.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/2c601ad9d2ff9bc8b282670cdd54f69f-Paper.pdf.

Saurav Muralidharan, Sharath Turuvekere Sreenivas, Raviraj Joshi, Marcin Chochowski, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, Jan Kautz, and Pavlo Molchanov. Compact language models via pruning and knowledge distillation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 41076–41102. Curran Associates, Inc., 2024. doi: 10.52202/079017-1299. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/4822991365c962105b1b95b1107d30e5-Paper-Conference.pdf.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Samet Oymak, Ankit Singh Rawat, Mahdi Soltanolkotabi, and Christos Thrampoulidis. On the role of attention in prompt-tuning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 26724–26768. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/oymak23a.html.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026. URL https://qwen.ai/blog?id=qwen3.8.

Oliver Sieberling, Denis Kuznedelev, Eldar Kurtic, and Dan Alistarh. EvoPress: Accurate dynamic model compression via evolutionary search. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 55556–55590. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/sieberling25a.html.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70): 1–57, 2018. URL http://jmlr.org/papers/v19/18-188.html.

Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. A simple and effective pruning approach for large language models, 2024. URL https://arxiv.org/abs/2306.11695.

Davoud Ataee Tarzanagh, Yingcong Li, Christos Thrampoulidis, and Samet Oymak. Transformers as support vector machines, 2024. URL https://arxiv.org/abs/2308.16898.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron Wallace, and David Bau. Function vectors in large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 17282–17333, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/4ae163cb8788970e53b4fd9578141139-Paper-Conference.pdf.

Bhavya Vasudeva, Puneesh Deora, and Christos Thrampoulidis. Implicit bias and fast convergence rates for self-attention, 2025. URL https://arxiv.org/abs/2402.05738.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5797–5808, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1580. URL https://aclanthology.org/P19-1580/.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small, 2022. URL https://arxiv.org/abs/2211.00593.

Kaiyue Wen, Yuchen Li, Bingbin Liu, and Andrej Risteski. Transformers are uninterpretable with myopic methods: a case study with bounded dyck grammars. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 38723–38766. Curran Associates, Inc., 2023. doi: 10.52202/075280-1681. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/79ba1b827d3fc58e129d1cbfc8ff69f2-Paper-Conference.pdf.

Diyuan Wu, Aleksandr Shevchenko, Samet Oymak, and Marco Mondelli. Attention with trained embeddings provably selects important tokens. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 28956–28997. Curran Associates, Inc., 2025. doi: 10.52202/085713-0972. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/29be2340edb8b224728248d4bb1ea9d4-Paper-Conference.pdf.

Mengzhou Xia, Zexuan Zhong, and Danqi Chen. Structured pruning learns compact and accurate models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1513–1528, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.107. URL https://aclanthology.org/2022.acl-long.107/.

Mengzhou Xia, Tianyu Gao, Zhiyuan Zeng, and Danqi Chen. Sheared llama: Accelerating language model pre-training via structured pruning. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 5385–5409, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/160adf2dc118a920e7858484b92a37d8-Paper-Conference.pdf.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Zhiguo Yang, Changjian Deng, Qinke Chen, Zijing Zhou, and Jian Cheng. Lsa: Layer-wise sparsity allocation for large language model pruning based on minimal linear reconstruction error. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 76241–76266, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/7b805585c7e249c1f65737506d4fe1e4-Paper-Conference.pdf.

Yunzhi Yao, Ningyu Zhang, Zekun Xi, Mengru Wang, Ziwen Xu, Shumin Deng, and Huajun Chen. Knowledge circuits in pretrained transformers. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 118571–118602. Curran Associates, Inc., 2024. doi: 10.52202/

079017-3765. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/d6df31b1be98e04be48af8bedb95b499-Paper-Conference.pdf.

Lu Yin, You Wu, Zhenyu Zhang, Cheng-Yu Hsieh, Yaqing Wang, Yiling Jia, Gen Li, Ajay Kumar Jaiswal, Mykola Pechenizkiy, Yi Liang, Michael Bendersky, Zhangyang Wang, and Shiwei Liu. Outlier weighed layerwise sparsity (OWL): A missing secret sauce for pruning LLMs to high sparsity. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57101–57115. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/yin24e.html.

A Details of Inter-Layer Budget Allocation

Let $\pi_\ell(j)$ denote the j-th head in layer ℓ sorted by ascending ACM score. Let $N_\ell(j)$ denote the per-token NLL increase caused by pruning only the first j candidate heads in layer ℓ, and let $Q_\ell(j)$ denote the top-1 prediction disagreement rate of the pruned model relative to the full unpruned model. Setting $N_\ell(0) = Q_\ell(0) = 0$, the raw marginal changes between adjacent prefixes are given by $N_\ell(j) - N_\ell(j-1)$ and $Q_\ell(j) - Q_\ell(j-1)$, respectively. Each marginal sequence is independently fitted using unit-weight least-squares isotonic regression, preserving negative fitted values and yielding $\widehat{\Delta}^{\mathrm{NLL}}_{\ell,j}$ and $\widehat{\Delta}^{\mathrm{pred}}_{\ell,j}$. While evaluating candidate prefixes in a given layer, all remaining layers are maintained in their unpruned model states.

The default probing subset comprises the final eight calibration blocks, using all available blocks when fewer than eight are present. Reference predictions from the unpruned model are evaluated once and shared across all probing steps. For each metric, we collect the absolute values of fitted marginal changes across all layers that exceed 10^{-12}; the scale factor is their 75th percentile, or one if no such values exist. The weights assigned to both normalized predictive metrics and the concentration penalty coefficient λ are set to one. Ties in raw ACM scores are broken by head index, and ties in allocation costs are resolved sequentially by layer index and candidate prefix order, ensuring that head selections within each layer strictly preserve the prefix structure.

Combining both predictive metrics with an intra-layer concentration penalty, the marginal cost of the j-th candidate is defined as

$$d_{\ell,j} = \frac{\widehat{\Delta}^{\mathrm{NLL}}_{\ell,j}}{s_{\mathrm{NLL}}} + \frac{\widehat{\Delta}^{\mathrm{pred}}_{\ell,j}}{s_{\mathrm{pred}}} + \lambda \frac{j-1}{\max(H_\ell - 1, 1)}. \tag{9}$$

Scale factors $s_{\mathrm{NLL}}, s_{\mathrm{pred}} > 0$ normalize the two predictive changes, while $\lambda \geq 0$ controls the concentration penalty that increases with the number of heads pruned in that layer.

The separable surrogate objective is formulated as

$$\min_{\substack{\sum_\ell k_\ell = K \\ 0 \leq k_\ell \leq H_\ell,\ k_\ell \in \mathbb{Z}}} \sum_\ell \left[\sum_{j=1}^{k_\ell} \left(\frac{\widehat{\Delta}^{\mathrm{NLL}}_{\ell,j}}{s_{\mathrm{NLL}}} + \frac{\widehat{\Delta}^{\mathrm{pred}}_{\ell,j}}{s_{\mathrm{pred}}} \right) + \lambda \frac{k_\ell(k_\ell - 1)}{2\max(H_\ell - 1, 1)} \right]. \tag{10}$$

Within each layer, isotonic regression makes both fitted marginal sequences non-decreasing. Since their scale factors are positive and $\lambda \geq 0$, the costs $d_{\ell,j}$ are non-decreasing in j, even when some fitted marginal changes are negative. Sorting all costs by value, then by layer index and candidate prefix order, makes the first K costs a feasible set of layer prefixes. Every feasible allocation selects K costs, whose sum cannot be smaller than the sum of these K lowest costs. Thus, the greedy allocation minimizes Eq. equation 10 for the fixed intra-layer ACM rankings. This result applies to the separable surrogate built from single-layer probes; it does not guarantee minimal NLL or prediction disagreement when heads from multiple layers are pruned together, since their effects can interact.

Table 2: Summary of controlled experiments on margin disparities across models and pruning ratios. Each cell reports the number of configurations where the low-margin group incurs greater damage than the comparison group over the total count (denominator is 15 in all cases). The PPL column indicates the count where the low-margin group exhibits a higher PPL ratio than the comparison group, and the Agr column indicates the count where the low-margin group achieves a lower top-1 agreement.

		LM vs HM		LM vs RD		ALM vs AHM		ALM vs ARD	
Model	Ratio	PPL	Agr	PPL	Agr	PPL	Agr	PPL	Agr
Llama-3.2-3B	5%	15	15	10	9	15	15	14	14
	10%	5	11	9	10	15	15	7	9
	20%	14	15	8	10	10	10	7	8
Qwen3-1.7B	5%	15	0	15	15	15	15	15	13
	10%	0	0	9	11	5	0	15	13
	20%	0	0	7	6	0	0	5	5
Qwen3-14B	5%	10	15	15	15	15	0	15	15
	10%	14	15	15	15	11	10	14	15
	20%	0	4	13	15	6	1	14	15

The probing stage evaluates the unpruned reference once and each of the H candidate prefixes once on the probing subset, requiring $H+1$ model evaluations over that subset. This cost is separate from the calibration pass used to collect ACM scoring statistics. The final pruned set is defined as

$$\mathcal{P}_r = \bigcup_{\ell} \{(\ell, \pi_\ell(j)) : 1 \le j \le k_\ell\}. \tag{11}$$

B Controlled Experiments on Margin Disparities

Experimental Design To examine whether the token competition margin correlates with the predictive contribution of attention heads to domain tasks, we conduct layer-wise count-matched controlled interventions on Llama-3.2-3B, Qwen3-1.7B, and Qwen3-14B across PetroBench, DISKOS-QA, and FormationEval. For each model-dataset pair, we use five calibration seeds (42, 56466, 153176, 65461, 25464), where each seed independently samples calibration blocks to compute head-level boundary scores and pre-projection head output magnitudes. Interventions are evaluated at three pruning ratios (5%, 10%, and 20%), comprising six intervention groups per setting: the low-margin group (LOW_MARGIN, selected in descending order of boundary score), the high-margin group (HIGH_MARGIN, selected in ascending order of boundary score), the count-matched random group (RANDOM), and three corresponding groups selected within two strata partitioned by the median pre-projection head output magnitude within each layer (ACT_MATCH_LOW_MARGIN, ACT_MATCH_HIGH_MARGIN, and ACT_MATCH_RANDOM). Magnitude stratification controls count matching within each layer-wise stratum, though it does not guarantee exact matching of mean magnitudes. Each intervention zeros out the output vectors of the selected heads while keeping all other layers unpruned, recording the resulting PPL ratio and top-1 agreement. In total, this protocol comprises 3 models $\times$ 3 datasets $\times$ 5 seeds $\times$ 3 pruning ratios = 135 comparison settings and 810 interventions.

Per-Model Results Table 2 reports the frequency with which the low-margin group incurs greater predictive damage than the comparison groups across models and pruning ratios. The denominator for each cell is 15 (3 datasets $\times$ 5 seeds).

Llama-3.2-3B provides consistent support for the low-margin hypothesis. The low-margin versus high-margin comparison (LM vs. HM) causes greater degradation in 14/15 or 15/15 cases at 5% and 20% pruning; while PPL reverses at 10% (5/15), top-1 agreement still shows greater damage in 11 of 15 cases. After magnitude stratification (ALM vs. AHM), the low-margin group degrades

performance in 40 of 45 cases, confirming that the margin signal persists after controlling for output magnitude.

For Qwen3-1.7B, at 5% pruning, LM vs. HM yields a higher PPL ratio in 15/15 cases and a lower top-1 agreement in 0/15 cases; at 10% and 20%, the count is 0/15 on both metrics. In this model (28 layers, 448 heads total), pruning 10% and 20% removes 45 and 90 heads, and removing high-margin heads at these rates causes a severe perplexity surge. Meanwhile, the low-margin group inflicts worse PPL than random pruning (LM vs. RD) in 31 of 45 cases.

For Qwen3-14B, LM vs. HM shows greater damage in 10/15 and 14/15 cases at 5% and 10% pruning, respectively, before reversing at 20% (0/15). This reversal is moderate (median high-margin PPL ratios range from 1.85 to 3.27) without catastrophic failure. Across all pruning ratios, LM vs. RD results in worse PPL in 43 of 45 cases and worse agreement in all 45 cases, exhibiting the strongest baseline signal among the three models.

Interpretation of High-Margin Group Reversals At higher pruning ratios in certain models, removing the high-margin group also incurs substantial predictive damage, so ACM does not rank heads by competition margin alone. In score component ablations, the composite score outperforms each individual term in both overall PPL ratio and top1 agreement.

C Complete Results on In-Domain Language Modeling and Prediction Fidelity

Figures 5 and 6 present comprehensive single-metric evaluation curves for seven models across Paloma-Dolma and the three domain datasets. The evaluated architectures encompass Llama-3.2-1B, Llama-3.2-3B, Llama-3.1-8B, Qwen3-1.7B, Qwen3-8B, Qwen3-14B, and Qwen3.8-27B; the comparison includes Random, Head Confidence, FLAP, ACM*, and ACM. Each method is evaluated across four pruning ratios: 5%, 10%, 20%, and 30%. ACM* selects heads globally based on model-wide score rankings, whereas ACM incorporates inter-layer budget allocation on top of the same scoring metric.

Across the 72 configurations comprising the first six models, Paloma-Dolma, DISKOS-QA, FormationEval, and four pruning ratios, ACM achieves a lower PPL ratio than ACM* in all 72 settings and a higher top-1 agreement in 48 settings. Among the 12 Qwen3.8-27B extension configurations, the corresponding counts are 12 and 9; at 30% pruning, ACM outperforms ACM* on both metrics in all three configurations. The agreement advantage is less consistent on Qwen3-8B and Qwen3-14B, and these curve intersections are preserved in the figures. The two figures also illustrate the full variation range of external baselines under higher pruning ratios.

To make every plotted operating point auditable, Tables 3–7 give the PPL ratio and top-1 agreement behind Figures 5 and 6. The Paloma-Dolma, DISKOS-QA, and FormationEval tables report PPL ratio followed by agreement in each method cell; the PetroBench results are separated by metric for readability. Values are rounded for display; all comparisons in the text use unrounded values. Calibration and probing blocks are drawn from the training split of each corpus (the official validation split for Paloma-Dolma), whereas PPL ratio and top-1 agreement are computed on the separate test split. The twelve model–dataset combinations in the main figure retain their original data aggregation. The remaining sixteen combinations provide supplementary plotted operating points; their calibration configurations are held fixed across methods.

D Complete Downstream Accuracy

Table 8 extends Table 1 to the Llama family. The Qwen3-8B and Qwen3-14B rows coincide with Table 1. Llama-3.2-1B, Llama-3.2-3B, and Llama-3.1-8B are evaluated at 20% and 30% pruning ratios, and Llama-3.2-1B is evaluated on PetroBench only. The unpruned accuracy of Llama-3.2-1B on PetroBench is 0.335, close to the 0.25 chance level of the label-balanced four-option subset, so differences among methods on this model are small.

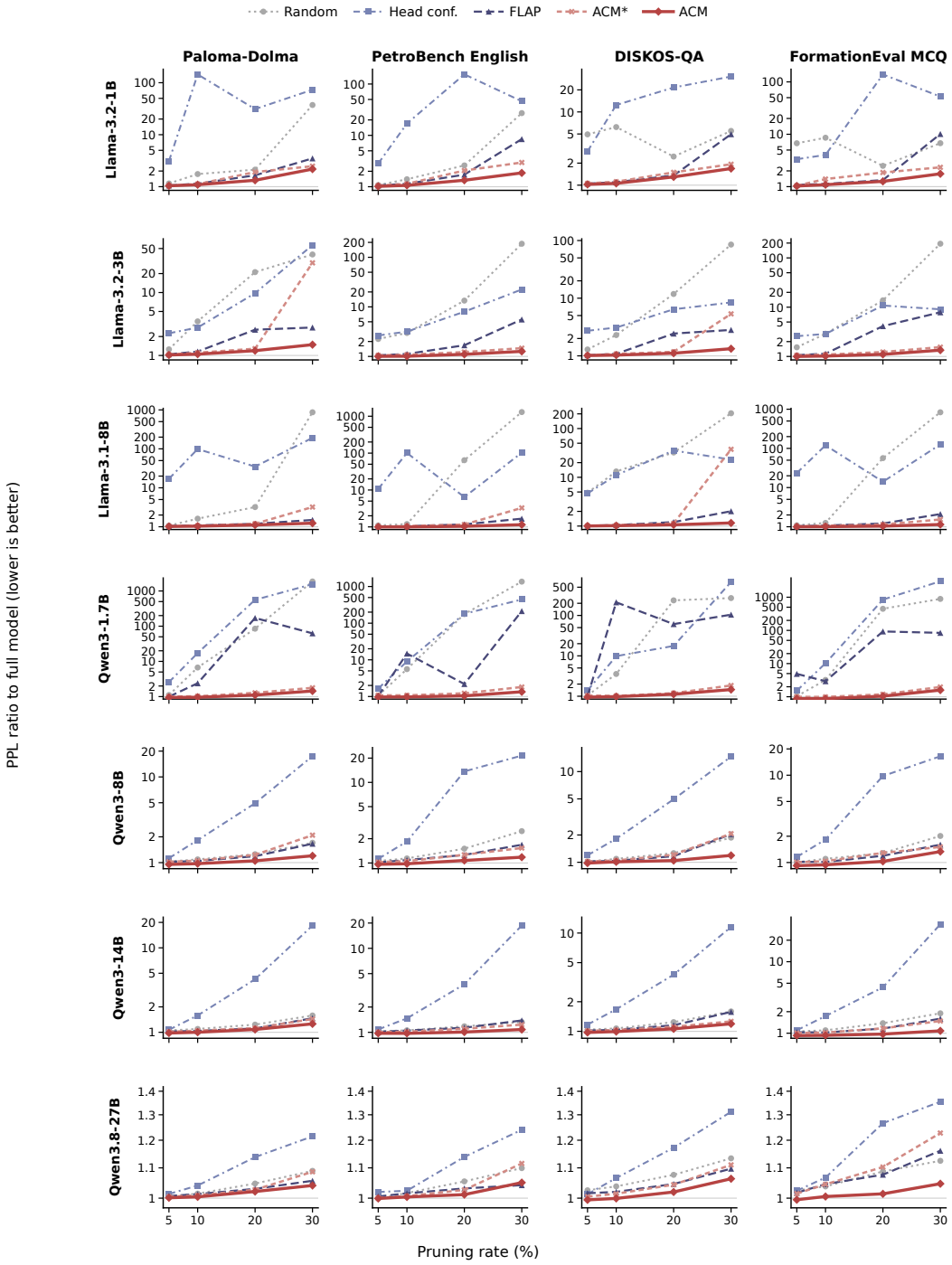

Figure 5: PPL ratio of seven models across Paloma-Dolma and three domain datasets. Each subplot compares five methods across four pruning ratios, where lower values indicate better performance. The vertical axis uses a logarithmic scale with independently calibrated ranges across subplots to display extreme values.

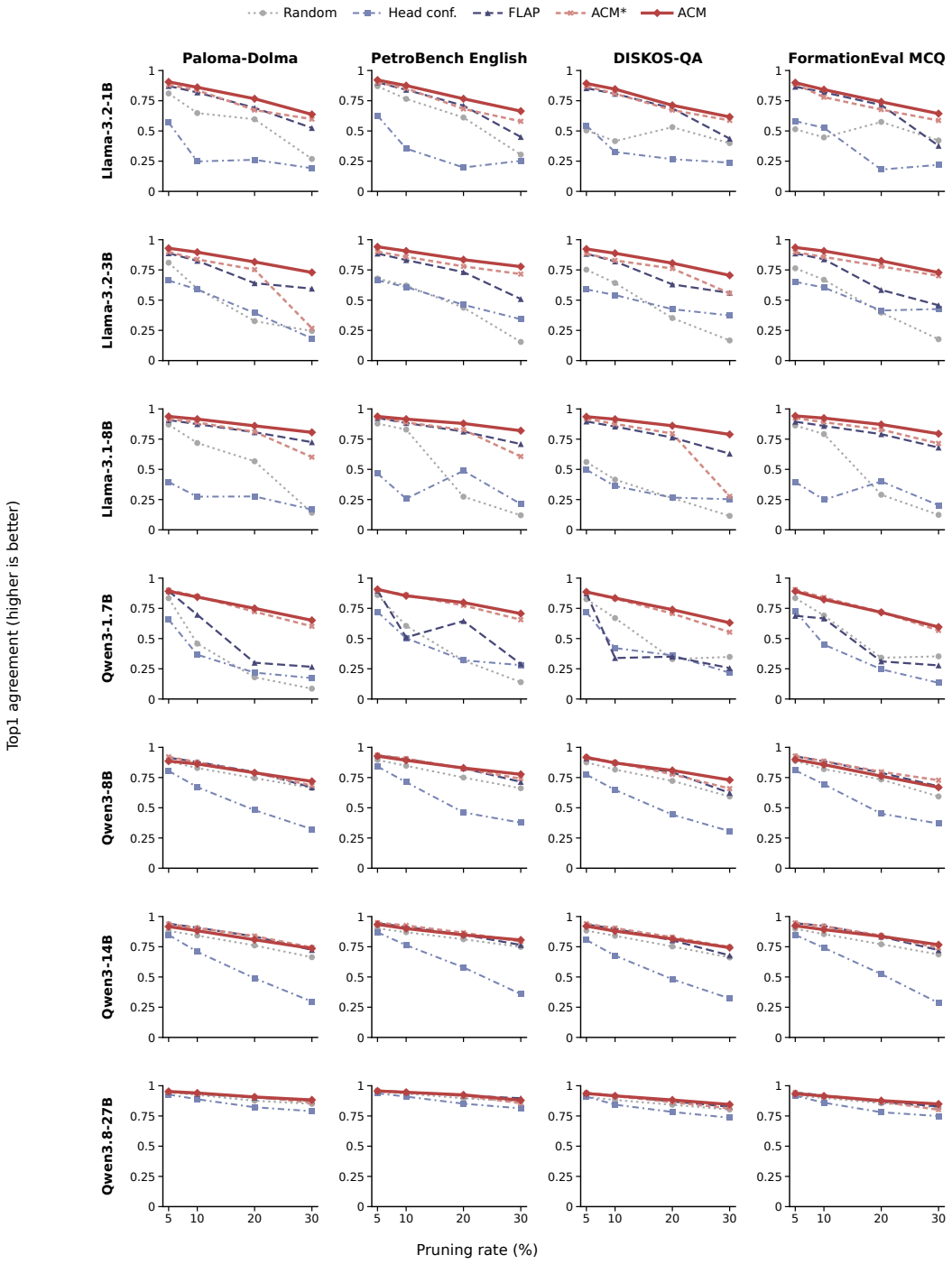

Figure 6: Top-1 agreement of seven models across Paloma-Dolma and three domain datasets. Methods, pruning ratios, and data sources match Figure 5. Higher values indicate better performance; all subplots share a common vertical axis from 0 to 1.

Table 3: Complete numerical results on Paloma-Dolma. Each method cell reports PPL ratio / top-1 agreement, with lower PPL ratio and higher agreement preferred. Displayed values are rounded from the plotted data.

Model	Pruned	Random	HeadConf	FLAP	ACM*	ACM
Llama-3.2-1B	5%	1.152/0.8103	3.096/0.5703	1.055/0.8719	1.053/0.8814	1.038/0.9059
Llama-3.2-1B	10%	1.748/0.6478	144.459/0.2479	1.143/0.8185	1.120/0.8368	1.090/0.8608
Llama-3.2-1B	20%	2.125/0.5965	31.108/0.2613	1.654/0.6917	1.912/0.6730	1.327/0.7664
Llama-3.2-1B	30%	37.258/0.2683	73.179/0.1918	3.500/0.5245	2.476/0.5980	2.179/0.6369
Llama-3.2-3B	5%	1.252/0.8100	2.253/0.6625	1.043/0.8881	1.035/0.8997	1.018/0.9293
Llama-3.2-3B	10%	3.481/0.5949	2.776/0.5885	1.155/0.8255	1.107/0.8375	1.050/0.8968
Llama-3.2-3B	20%	21.094/0.3265	9.752/0.3942	2.567/0.6406	1.281/0.7529	1.192/0.8163
Llama-3.2-3B	30%	40.231/0.2447	56.676/0.1839	2.781/0.5953	29.608/0.2675	1.485/0.7292
Llama-3.1-8B	5%	1.068/0.8703	17.009/0.3955	1.043/0.9080	1.033/0.9152	1.018/0.9376
Llama-3.1-8B	10%	1.612/0.7182	98.862/0.2738	1.084/0.8741	1.059/0.8873	1.037/0.9157
Llama-3.1-8B	20%	3.200/0.5668	35.034/0.2760	1.192/0.8069	1.195/0.8103	1.113/0.8601
Llama-3.1-8B	30%	869.055/0.1405	191.393/0.1672	1.496/0.7249	3.207/0.6013	1.234/0.8061
Qwen3-1.7B	5%	1.109/0.8335	2.577/0.6568	0.995/0.8927	1.007/0.9012	0.947/0.8914
Qwen3-1.7B	10%	6.701/0.4579	17.244/0.3667	2.396/0.6969	1.024/0.8498	0.958/0.8434
Qwen3-1.7B	20%	84.896/0.1797	560.679/0.2184	170.984/0.3001	1.281/0.7225	1.106/0.7497
Qwen3-1.7B	30%	1860.835/0.0855	1560.588/0.1732	62.652/0.2662	1.775/0.6014	1.441/0.6508
Qwen3-8B	5%	1.031/0.8834	1.127/0.8051	1.023/0.9141	1.022/0.9215	0.953/0.8862
Qwen3-8B	10%	1.091/0.8288	1.829/0.6713	1.044/0.8793	1.055/0.8769	0.973/0.8624
Qwen3-8B	20%	1.253/0.7457	4.924/0.4801	1.188/0.7958	1.232/0.7891	1.050/0.7894
Qwen3-8B	30%	1.704/0.6641	17.485/0.3231	1.662/0.6689	2.090/0.6837	1.201/0.7187
Qwen3-14B	5%	1.053/0.8814	1.071/0.8486	1.021/0.9407	1.016/0.9379	0.980/0.9169
Qwen3-14B	10%	1.095/0.8430	1.582/0.7084	1.040/0.9082	1.045/0.9052	1.001/0.8822
Qwen3-14B	20%	1.233/0.7608	4.248/0.4892	1.110/0.8354	1.121/0.8386	1.080/0.8082
Qwen3-14B	30%	1.581/0.6635	18.345/0.2964	1.467/0.7235	1.435/0.7439	1.259/0.7356
Qwen3.8-27B	5%	1.006/0.9467	1.013/0.9275	1.007/0.9497	1.004/0.9501	1.001/0.9528
Qwen3.8-27B	10%	1.014/0.9241	1.041/0.8882	1.013/0.9337	1.009/0.9354	1.004/0.9400
Qwen3.8-27B	20%	1.047/0.8760	1.138/0.8222	1.030/0.9020	1.027/0.9071	1.021/0.9071
Qwen3.8-27B	30%	1.090/0.8500	1.215/0.7896	1.056/0.8767	1.087/0.8605	1.041/0.8829

Table 4: Complete numerical results on DISKOS-QA. Each method cell reports PPL ratio / top-1 agreement, with lower PPL ratio and higher agreement preferred. Displayed values are rounded from the plotted data.

Model	Pruned	Random	HeadConf	FLAP	ACM*	ACM
Llama-3.2-1B	5%	4.940/0.5039	2.896/0.5445	1.051/0.8538	1.058/0.8692	1.024/0.8928
Llama-3.2-1B	10%	6.212/0.4158	12.419/0.3261	1.099/0.8074	1.115/0.8082	1.058/0.8457
Llama-3.2-1B	20%	2.459/0.5312	21.555/0.2663	1.356/0.6871	1.495/0.6722	1.296/0.7123
Llama-3.2-1B	30%	5.502/0.3989	30.416/0.2377	4.960/0.4364	1.931/0.5874	1.686/0.6163
Llama-3.2-3B	5%	1.290/0.7520	2.721/0.5890	1.026/0.8833	1.028/0.8862	1.012/0.9231
Llama-3.2-3B	10%	2.294/0.6418	3.095/0.5421	1.082/0.8206	1.081/0.8284	1.031/0.8878
Llama-3.2-3B	20%	11.834/0.3520	6.417/0.4254	2.435/0.6304	1.175/0.7617	1.111/0.8060
Llama-3.2-3B	30%	85.124/0.1665	8.444/0.3730	2.816/0.5597	5.348/0.5566	1.330/0.7053
Llama-3.1-8B	5%	4.821/0.5609	4.668/0.4971	1.017/0.8962	1.013/0.9189	1.005/0.9354
Llama-3.1-8B	10%	13.190/0.4149	11.087/0.3614	1.049/0.8539	1.038/0.8757	1.019/0.9150
Llama-3.1-8B	20%	32.276/0.2623	35.474/0.2662	1.208/0.7634	1.137/0.7963	1.063/0.8608
Llama-3.1-8B	30%	205.830/0.1149	23.025/0.2528	2.015/0.6297	37.496/0.2761	1.158/0.7891
Qwen3-1.7B	5%	1.040/0.8257	1.388/0.7212	0.996/0.8742	1.046/0.8846	0.950/0.8857
Qwen3-1.7B	10%	3.559/0.6695	9.841/0.4213	211.297/0.3390	1.019/0.8311	0.979/0.8348
Qwen3-1.7B	20%	236.320/0.3301	17.588/0.3612	61.050/0.3505	1.191/0.7071	1.120/0.7392
Qwen3-1.7B	30%	269.171/0.3483	672.576/0.2172	104.491/0.2560	1.846/0.5528	1.462/0.6306
Qwen3-8B	5%	1.019/0.8758	1.194/0.7741	1.020/0.9087	1.023/0.9191	0.980/0.9169
Qwen3-8B	10%	1.091/0.8158	1.818/0.6478	1.038/0.8745	1.046/0.8760	1.011/0.8707
Qwen3-8B	20%	1.261/0.7221	4.949/0.4430	1.166/0.7928	1.212/0.7765	1.044/0.8087
Qwen3-8B	30%	1.870/0.5934	14.632/0.3061	2.043/0.6231	2.067/0.6587	1.190/0.7296
Qwen3-14B	5%	1.040/0.8835	1.170/0.8058	1.018/0.9340	1.011/0.9403	0.974/0.9204
Qwen3-14B	10%	1.070/0.8407	1.675/0.6770	1.037/0.9001	1.031/0.9040	0.994/0.8796
Qwen3-14B	20%	1.240/0.7523	3.766/0.4810	1.152/0.8026	1.112/0.8294	1.060/0.8148
Qwen3-14B	30%	1.592/0.6616	11.530/0.3222	1.576/0.6784	1.261/0.7475	1.192/0.7425
Qwen3.8-27B	5%	1.025/0.9066	1.013/0.9110	1.016/0.9389	1.005/0.9406	0.995/0.9361
Qwen3.8-27B	10%	1.038/0.8827	1.065/0.8434	1.020/0.9200	1.014/0.9145	0.999/0.9162
Qwen3.8-27B	20%	1.076/0.8424	1.172/0.7834	1.046/0.8699	1.043/0.8638	1.020/0.8823
Qwen3.8-27B	30%	1.133/0.8027	1.312/0.7352	1.097/0.8264	1.110/0.8126	1.063/0.8450

Table 5: Complete numerical results on FormationEval MCQ. Each method cell reports PPL ratio / top-1 agreement, with lower PPL ratio and higher agreement preferred. Displayed values are rounded from the plotted data.

Model	Pruned	Random	HeadConf	FLAP	ACM*	ACM
Llama-3.2-1B	5%	6.735/0.5145	3.337/0.5812	1.081/0.8658	1.044/0.8888	1.028/0.8992
Llama-3.2-1B	10%	8.590/0.4464	4.032/0.5259	1.127/0.8198	1.397/0.7793	1.089/0.8419
Llama-3.2-1B	20%	2.495/0.5745	138.638/0.1805	1.335/0.7149	1.860/0.6768	1.261/0.7406
Llama-3.2-1B	30%	6.754/0.4213	52.515/0.2194	10.115/0.3775	2.336/0.5865	1.759/0.6432
Llama-3.2-3B	5%	1.534/0.7637	2.577/0.6513	1.064/0.8865	1.056/0.8971	1.001/0.9357
Llama-3.2-3B	10%	2.889/0.6684	2.867/0.6043	1.137/0.8382	1.091/0.8575	1.019/0.9067
Llama-3.2-3B	20%	13.762/0.3969	10.804/0.4144	4.125/0.5845	1.230/0.7796	1.102/0.8247
Llama-3.2-3B	30%	194.010/0.1770	9.056/0.4265	7.862/0.4578	1.545/0.7012	1.341/0.7283
Llama-3.1-8B	5%	1.089/0.8628	23.550/0.3955	1.035/0.8941	1.027/0.9237	0.997/0.9420
Llama-3.1-8B	10%	1.253/0.7918	119.794/0.2485	1.075/0.8595	1.061/0.8904	1.007/0.9240
Llama-3.1-8B	20%	56.476/0.2899	14.083/0.4003	1.211/0.7911	1.142/0.8289	1.039/0.8711
Llama-3.1-8B	30%	834.462/0.1234	126.297/0.2029	2.107/0.6804	1.530/0.7139	1.141/0.7946
Qwen3-1.7B	5%	1.035/0.8346	1.518/0.7268	4.892/0.6882	0.995/0.9046	0.877/0.8911
Qwen3-1.7B	10%	3.203/0.6908	10.229/0.4500	2.938/0.6664	0.988/0.8386	0.875/0.8232
Qwen3-1.7B	20%	447.955/0.3414	846.098/0.2457	92.930/0.3111	1.185/0.7193	1.044/0.7168
Qwen3-1.7B	30%	888.656/0.3529	3022.707/0.1338	84.306/0.2781	1.967/0.5674	1.586/0.5957
Qwen3-8B	5%	1.009/0.8887	1.158/0.8091	1.018/0.9258	1.007/0.9283	0.920/0.8986
Qwen3-8B	10%	1.102/0.8195	1.832/0.6942	1.016/0.8843	1.036/0.8858	0.941/0.8568
Qwen3-8B	20%	1.282/0.7344	9.787/0.4512	1.194/0.7894	1.288/0.7976	1.030/0.7619
Qwen3-8B	30%	2.013/0.5934	16.473/0.3694	1.601/0.6775	1.508/0.7278	1.333/0.6694
Qwen3-14B	5%	1.060/0.8985	1.099/0.8481	1.033/0.9444	1.006/0.9471	0.927/0.9236
Qwen3-14B	10%	1.098/0.8544	1.737/0.7410	1.025/0.9229	1.022/0.9203	0.934/0.8917
Qwen3-14B	20%	1.381/0.7706	4.383/0.5241	1.161/0.8340	1.169/0.8429	0.972/0.8361
Qwen3-14B	30%	1.897/0.6868	33.396/0.2848	1.598/0.7221	1.482/0.7400	1.079/0.7667
Qwen3.8-27B	5%	1.025/0.9177	1.024/0.9177	1.017/0.9444	1.014/0.9438	0.995/0.9338
Qwen3.8-27B	10%	1.038/0.8985	1.067/0.8596	1.045/0.9193	1.044/0.9114	1.005/0.9154
Qwen3.8-27B	20%	1.088/0.8555	1.266/0.7809	1.077/0.8682	1.103/0.8669	1.014/0.8772
Qwen3.8-27B	30%	1.125/0.8228	1.354/0.7490	1.161/0.8291	1.227/0.8022	1.046/0.8489

Table 6: Complete numerical results on PetroBench English: PPL ratio. Lower is better. Displayed values are rounded from the plotted data.

Model	Pruned	Random	HeadConf	FLAP	ACM*	ACM
Llama-3.2-1B	5%	1.078	2.827	1.060	1.052	1.016
Llama-3.2-1B	10%	1.398	16.998	1.140	1.140	1.064
Llama-3.2-1B	20%	2.598	154.266	1.718	2.065	1.336
Llama-3.2-1B	30%	27.124	46.162	8.455	2.963	1.857
Llama-3.2-3B	5%	2.249	2.646	1.053	1.049	1.005
Llama-3.2-3B	10%	2.975	3.165	1.122	1.084	1.010
Llama-3.2-3B	20%	13.379	7.987	1.681	1.224	1.103
Llama-3.2-3B	30%	187.746	22.467	5.564	1.468	1.264
Llama-3.1-8B	5%	1.097	10.704	1.018	1.022	0.999
Llama-3.1-8B	10%	1.193	101.007	1.043	1.057	1.001
Llama-3.1-8B	20%	64.232	6.436	1.172	1.159	1.041
Llama-3.1-8B	30%	1295.538	102.623	1.670	3.274	1.149
Qwen3-1.7B	5%	1.062	1.647	1.024	1.080	0.969
Qwen3-1.7B	10%	5.579	9.031	14.817	1.078	0.974
Qwen3-1.7B	20%	176.899	180.203	2.190	1.222	1.043
Qwen3-1.7B	30%	1357.565	442.458	213.679	1.815	1.343
Qwen3-8B	5%	1.061	1.126	1.018	1.028	0.962
Qwen3-8B	10%	1.139	1.848	1.087	1.053	0.978
Qwen3-8B	20%	1.505	13.680	1.259	1.265	1.074
Qwen3-8B	30%	2.484	21.485	1.683	1.533	1.180
Qwen3-14B	5%	1.030	1.083	1.015	1.007	0.973
Qwen3-14B	10%	1.056	1.467	1.050	1.018	0.972
Qwen3-14B	20%	1.179	3.731	1.133	1.103	1.008
Qwen3-14B	30%	1.326	18.684	1.385	1.235	1.080
Qwen3.8-27B	5%	0.999	1.020	1.005	1.001	0.999
Qwen3.8-27B	10%	1.016	1.024	1.015	1.006	1.004
Qwen3.8-27B	20%	1.054	1.139	1.031	1.026	1.011
Qwen3.8-27B	30%	1.099	1.240	1.042	1.115	1.050

Table 7: Complete numerical results on PetroBench English: top-1 agreement. Higher is better. Displayed values are rounded from the plotted data.

Model	Pruned	Random	HeadConf	FLAP	ACM*	ACM
Llama-3.2-1B	5%	0.8702	0.6268	0.9017	0.9019	0.9212
Llama-3.2-1B	10%	0.7651	0.3541	0.8407	0.8522	0.8753
Llama-3.2-1B	20%	0.6112	0.1978	0.7096	0.6827	0.7663
Llama-3.2-1B	30%	0.3040	0.2524	0.4502	0.5797	0.6642
Llama-3.2-3B	5%	0.6788	0.6635	0.8854	0.9013	0.9414
Llama-3.2-3B	10%	0.6207	0.6085	0.8316	0.8590	0.9069
Llama-3.2-3B	20%	0.4380	0.4615	0.7321	0.7794	0.8347
Llama-3.2-3B	30%	0.1544	0.3422	0.5095	0.7159	0.7774
Llama-3.1-8B	5%	0.8786	0.4656	0.9252	0.9323	0.9376
Llama-3.1-8B	10%	0.8294	0.2572	0.8837	0.8878	0.9157
Llama-3.1-8B	20%	0.2739	0.4881	0.8128	0.8269	0.8794
Llama-3.1-8B	30%	0.1200	0.2134	0.7097	0.6065	0.8199
Qwen3-1.7B	5%	0.8598	0.7208	0.8972	0.9015	0.9069
Qwen3-1.7B	10%	0.6047	0.5029	0.5100	0.8608	0.8539
Qwen3-1.7B	20%	0.3186	0.3201	0.6451	0.7739	0.7981
Qwen3-1.7B	30%	0.1393	0.2803	0.2901	0.6559	0.7068
Qwen3-8B	5%	0.8952	0.8414	0.9318	0.9361	0.9271
Qwen3-8B	10%	0.8473	0.7115	0.9031	0.9069	0.8924
Qwen3-8B	20%	0.7501	0.4612	0.8265	0.8236	0.8291
Qwen3-8B	30%	0.6614	0.3762	0.7144	0.7406	0.7768
Qwen3-14B	5%	0.9027	0.8696	0.9444	0.9489	0.9350
Qwen3-14B	10%	0.8703	0.7633	0.9146	0.9261	0.9004
Qwen3-14B	20%	0.8119	0.5792	0.8503	0.8657	0.8478
Qwen3-14B	30%	0.7484	0.3584	0.7648	0.7935	0.8059
Qwen3.8-27B	5%	0.9532	0.9372	0.9601	0.9597	0.9565
Qwen3.8-27B	10%	0.9352	0.9102	0.9463	0.9458	0.9460
Qwen3.8-27B	20%	0.8958	0.8518	0.9186	0.9158	0.9245
Qwen3.8-27B	30%	0.8706	0.8129	0.8966	0.8562	0.8801

Table 8: Complete downstream accuracy on the PetroBench English SCQ balanced evaluation subset and FormationEval full MCQ. Full denotes the unpruned model; both benchmarks employ the candidate label length-normalized likelihood scoring protocol.

Model	Pruned	Full	Random	HeadConf	FLAP	ACM
PetroBench English SCQ						
Qwen3-8B	5%	0.776	0.751	0.761	0.764	**0.766**
	10%	0.776	0.732	0.646	0.760	**0.766**
	20%	0.776	0.575	0.258	0.727	**0.737**
	30%	0.776	0.392	0.245	0.631	**0.687**
Qwen3-14B	5%	0.792	0.779	0.751	0.784	**0.798**
	10%	0.792	0.750	0.625	**0.789**	**0.789**
	20%	0.792	0.661	0.275	0.758	**0.780**
	30%	0.792	0.430	0.249	0.641	**0.710**
Llama-3.2-1B	20%	0.335	0.257	0.243	0.261	**0.293**
	30%	0.335	0.253	0.243	0.249	**0.274**
Llama-3.2-3B	20%	0.653	0.333	0.243	**0.550**	0.489
	30%	0.653	0.257	0.233	0.418	**0.436**
Llama-3.1-8B	20%	0.643	0.349	0.254	0.526	**0.562**
	30%	0.643	0.256	0.233	0.290	**0.497**
FormationEval full MCQ						
Qwen3-8B	5%	0.855	0.802	0.780	0.850	**0.859**
	10%	0.855	0.711	0.568	**0.836**	**0.836**
	20%	0.855	0.527	0.288	0.780	**0.813**
	30%	0.855	0.312	0.285	0.692	**0.727**
Qwen3-14B	5%	0.854	0.824	0.760	0.846	**0.859**
	10%	0.854	0.790	0.646	0.838	**0.850**
	20%	0.854	0.663	0.296	0.786	**0.833**
	30%	0.854	0.380	0.250	0.735	**0.808**
Llama-3.2-3B	20%	0.721	0.273	0.271	**0.470**	0.413
	30%	0.721	0.243	0.273	0.315	**0.330**
Llama-3.1-8B	20%	0.758	0.343	0.275	0.652	**0.660**
	30%	0.758	0.269	0.277	**0.564**	0.403

E Candidate Selection Consistency Across Calibration Subsets

For each model, we construct three non-overlapping subsets A, B, and C from the PetroBench calibration corpus. Each subset contains 128 calibration blocks of 256 tokens, starting at block offsets 0, 128, and 256, respectively; all scores use seed 42 and temperature $\tau = 0.5$. We compute ACM head scores separately on each subset, calculate pairwise Spearman correlations over the ranks of all heads, and compare the Jaccard similarities of the top 10% highest-scoring head sets.

The three highest-scoring sets share 104 of 115 heads for Qwen3-8B and 148 of 160 for Qwen3-14B, corresponding to 90.4% and 92.5% of each model's candidate set.

F Experimental Details of Calibration-Corpus Transfer

We use Qwen3-8B to compare attention-head masks obtained by calibrating on PetroBench petroleum-domain text and WikiText-2 general-domain text. The petroleum calibration corpus consists of PetroBench answer-and-analysis text, whereas the general-domain calibration corpus is the WikiText-2 training text.

For each calibration corpus, we use two non-overlapping windows. The first comprises the initial 256 blocks, and the second comprises 256 additional blocks beginning at block 2048; each block

Table 9: Consistency of ACM candidate selection across calibration subsets. Spearman correlations use the ranks of all heads; Jaccard similarities use the top 10% highest-scoring head sets. The intersection column reports the number of heads shared by a pair of sets and the size of each set. Means are arithmetic averages across the three subset pairs.

Model	Subset pair	Spearman	Intersection / set size	Jaccard
Qwen3-8B	A/B	0.9946	110/115	0.9167
	A/C	0.9842	104/115	0.8254
	B/C	0.9886	107/115	0.8699
	Mean	0.9891		0.8707
Qwen3-14B	A/B	0.9947	154/160	0.9277
	A/C	0.9853	150/160	0.8824
	B/C	0.9874	151/160	0.8935
	Mean	0.9891		0.9012

contains 256 tokens. At pruning ratios of 20% and 30%, we generate one mask from each calibration corpus and evaluate the resulting masks on independent test text using the per-token NLL increase relative to the unpruned model.

Each test set contains 64 equally spaced blocks. The horizontal axis in Figure 3(a) reports the difference between the per-token NLL increases of the PetroBench-calibrated and WikiText-2-calibrated masks. Both increases are measured relative to the unpruned model. Horizontal error bars are obtained from 20,000 bootstrap replicates, treating each group of four consecutive test blocks as one resampling unit. We report the corresponding 95% percentile interval. These intervals quantify only uncertainty from test-text sampling.

The mask-overlap ratio is defined as the number of heads pruned by both masks divided by the number of heads pruned by the corresponding individual mask. The layer- and head-wise masks for the first calibration window at a 20% pruning ratio are shown in Figure 3(b).

G Examples of Evaluation Data

The PetroBench English SCQ evaluation set contains 1,000 questions sampled with seed 42 from 19 English single-choice files containing 14,732 questions in total. The four answer labels are balanced at 250 questions each. FormationEval full MCQ contains all 505 questions in the official version used here. For both tasks, the model receives the question and all four labeled options, followed by an `Answer:` prefix. We score each candidate answer label by length-normalized negative log-likelihood and select the label with the lowest score. The gold label is used only to calculate accuracy.

G.1 PetroBench English SCQ

One evaluation item, `Q00515`, has the following input and gold answer. The displayed gold answer is excluded from the model input.

> Question: Which of the following descriptions about the fullness of oil and gas reservoirs is correct?
> A. The fill degree is equal to the ratio of the height of oil and gas to the thickness of the reservoir.
> B. The fill degree can be obtained from the ratio of the area of oil and gas to the area of the reservoir.
> C. The value of the fill degree is always less than 1.
> D. The fill degree is not affected by the physical properties of the reservoir.
> Gold answer: A.

G.2 FORMATIONEVAL MCQ

The following item, `formationeval_v0.1_petrophysics_gamma_interactions_006`, comes from the 505-question evaluation set (Ermilov, 2026). As with PetroBench, the gold answer is used only for accuracy calculation.

> Question: What is the minimum threshold energy required for a gamma ray to undergo pair production?
> A. 0.511 MeV
> B. 1.022 MeV
> C. 2.22 MeV
> D. 100 keV
> Gold answer: B.

G.3 DISKOS-QA

The following question and answer occur in the held-out DISKOS-QA text used for language-modeling evaluation (georgeghon & contributors, 2026). The source context that precedes them in the test document is omitted here for brevity; this corpus is evaluated with PPL ratio and next-token agreement, not multiple-choice accuracy.

> Question: What are the implications of a 1.14 g/cm3 Draupne gradient for hydrostatic depth estimation and overpressure trend?
> Answer: A 1.14 g/cm3 gradient indicates elevated pore pressure (overpressure) relative to normal hydrostatic, so hydrostatic depth is effectively shallower and an increasing overpressure trend exists toward the top reservoir.

G.4 PALOMA-DOLMA

Paloma-Dolma is likewise evaluated as text rather than as a question-answering task. The following excerpt is from the first document in the held-out `wiki` source (Magnusson et al., 2023); the evaluation aggregates five sources: `common-crawl`, `pes2o`, `reddit_uniform`, `stack_uniform`, and `wiki`.

> 61st Signal Regiment was from 1939-67 a communications regiment of the British Army, part of the Royal Signals.

H PHYSICAL PRUNING AND INPUT PROCESSING EFFICIENCY OF FULL ACM

To assess actual deployment gains, we physically prune Llama-3.2-3B on DISKOS-QA by directly executing the head-level selections and inter-layer budget allocations derived from full ACM at 20% and 30% query head sparsity. The unpruned model comprises 28 layers, each featuring 24 query heads and 8 grouped key-value heads; these target ratios correspond to the physical excision of 134 and 202 query heads, respectively. Following original head indices, we slice the corresponding rows from query projection matrices and columns from attention output projection matrices while leaving key-value projections intact, tracking surviving query-to-KV head associations through explicit indexing. Retained query heads vary across layers, whereas residual dimensions, head dimensions, and key-value cache head counts remain unchanged. This physical excision reduces overall model parameters by 3.28% and 4.94%; all subsequent mentions of 20% and 30% refer strictly to query head pruning ratios.

Our verification protocol inspects post-pruning weight tensor shapes, query-to-KV head alignments, full-sequence prefill execution, and autoregressive decoding with dynamic caching. FP64 regression checks on a miniature architecture and independent FP32 audits on the full Llama-3.2-3B model verify numerical equivalence between the physically excised model and an output-zeroed surrogate sharing identical retained heads. Any operating points exceeding numerical tolerances under BF16 are further audited in FP32. Quality benchmarking adopts the three calibration configurations, profiling 256 text chunks of 256 tokens each per configuration. Because evaluation windows overlap

Table 10: Head-level physical pruning quality and input processing speedups of Llama-3.2-3B on DISKOS-QA. PPL and top1 agreement report means and standard deviations across three calibration settings. Speedup column headers indicate batch size $\times$ input sequence length, and values > 1 denote speedup factors over the unpruned baseline. FLAP serves as a score-adapted baseline under the identical head-level physical excision, omitting the original bias compensation.

Query head pruning ratio	Method	PPL	Top-1 agreement (%)	4×1024	4×4096
20%	FLAP	67.006 ± 1.347	63.312 ± 0.331	1.023	1.033
	Full ACM	30.580 ± 0.171	81.117 ± 0.492	1.023	1.041
30%	FLAP	77.495 ± 1.081	56.461 ± 0.202	1.041	1.067
	Full ACM	36.594 ± 0.927	71.053 ± 1.258	1.044	1.073

Table 11: Paired wins of the composite ACM score over each scoring alternative. Each cell reports the number of settings in which ACM has both a lower PPL ratio and a higher top-1 agreement than the alternative. Each model-ratio row contains nine settings (three datasets and three calibration windows).

Model	Pruned	Margin only	Activation only	Multiplicative
Llama-3.2-3B	5%	6/9	9/9	9/9
	10%	7/9	9/9	9/9
	20%	6/9	9/9	9/9
	30%	8/9	9/9	8/9
	All	27/36	36/36	35/36
Qwen3-1.7B	5%	9/9	9/9	9/9
	10%	9/9	9/9	9/9
	20%	9/9	9/9	9/9
	30%	9/9	9/9	9/9
	All	36/36	36/36	36/36
Qwen3-14B	5%	6/9	9/9	9/9
	10%	6/9	9/9	9/9
	20%	7/9	9/9	9/9
	30%	9/9	9/9	9/9
	All	28/36	36/36	36/36
Qwen3.8-27B	5%	9/9	4/9	9/9
	10%	9/9	6/9	9/9
	20%	9/9	7/9	9/9
	30%	9/9	7/9	8/9
	All	36/36	24/36	35/36
All models	All	127/144	132/144	142/144

across configurations, the aggregated means and standard deviations in Table 10 do not denote statistical significance over mutually disjoint test sets.

I Formula Component Ablation by Model and Pruning Ratio

Table 11 breaks down the four scoring formulas by model and pruning ratio. All variants use global top-K head selection under the same pruning budget; the composite ACM score therefore follows the ACM* selection protocol in Section 4.1. Each comparison is paired by model, dataset, calibration window, and pruning ratio.

J Complete Evaluation of Temperature Sensitivity

We evaluate Llama-3.2-3B, Qwen3-1.7B, and Qwen3-14B on PetroBench, DISKOS-QA, and FormationEval across four pruning ratios (5%, 10%, 20%, and 30%), totaling 36 experimental settings. Each setting evaluates seven candidate temperatures, $\tau \in \{0.125, 0.25, 0.375, 0.5, 0.75, 1, 2\}$,

Table 12: Summary of performance gaps for seven candidate temperatures across 36 model, dataset, and pruning ratio settings. Reference values for each setting are determined by the optimal three-window mean across the seven candidate temperatures. Lower PPL gaps and lower top-1 agreement gaps indicate better performance; the former is reported as a percentage, and the latter in percentage points.

Temperature τ	Mean PPL gap	Max PPL gap	Mean agreement gap	Max agreement gap
0.125	0.706	4.634	0.371	1.304
0.25	0.890	7.976	0.437	2.153
0.375	0.547	2.675	0.296	1.173
0.5	0.441	1.883	0.229	1.188
0.75	1.050	4.117	0.519	1.755
1	1.455	12.864	0.727	4.286
2	1.520	12.274	0.948	4.560

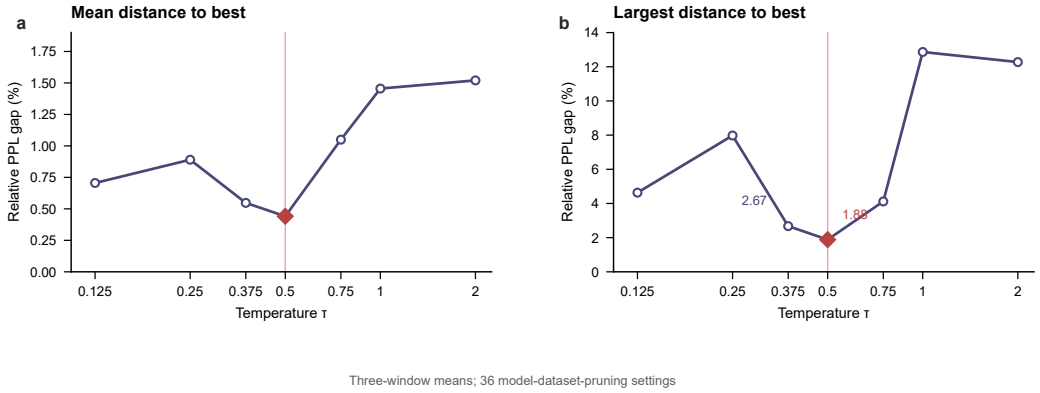

Figure 7: Relative PPL gaps of seven candidate temperatures across 36 settings. The left panel shows the average gap across settings, and the right panel shows the maximum gap. Each setting first averages across three calibration windows before comparing against the minimum mean among the seven temperatures. Red diamonds and vertical lines mark the default temperature $\tau = 0.5$.

across three calibration windows corresponding to seeds 42, 43, and 44, reporting the mean across windows. Because the three calibration windows overlap, the reported means and gaps characterize sensitivity across temperatures rather than serving as statistical significance tests over independent trials.

For each setting, we compute the mean PPL ratio and top-1 agreement across the three windows, adopting the minimum mean PPL ratio and maximum mean top-1 agreement across the seven temperatures as the optimal reference values. The relative PPL gap is defined as the current mean divided by the minimum mean minus one, reported as a percentage; the top-1 agreement gap is defined as the maximum mean minus the current mean, reported in percentage points. Table 12 summarizes the average and maximum gaps across the 36 settings. Fixing $\tau = 0.5$ yields an average relative PPL gap of 0.441% and a maximum gap of 1.883%, both lower than those of all other candidate temperatures; its average top-1 agreement gap is 0.229 percentage points, with a maximum gap of 1.188 percentage points.

Figure 7 illustrates the mean and maximum relative PPL gaps across the 36 settings for each candidate temperature. Both summary metrics attain their minimum simultaneously at $\tau = 0.5$, although the optimal temperature varies across individual settings. Figure 8 further details the distribution of relative PPL gaps across all individual settings; even when aggregate gaps at a fixed temperature remain modest, optimal temperatures vary across models, datasets, and pruning ratios.

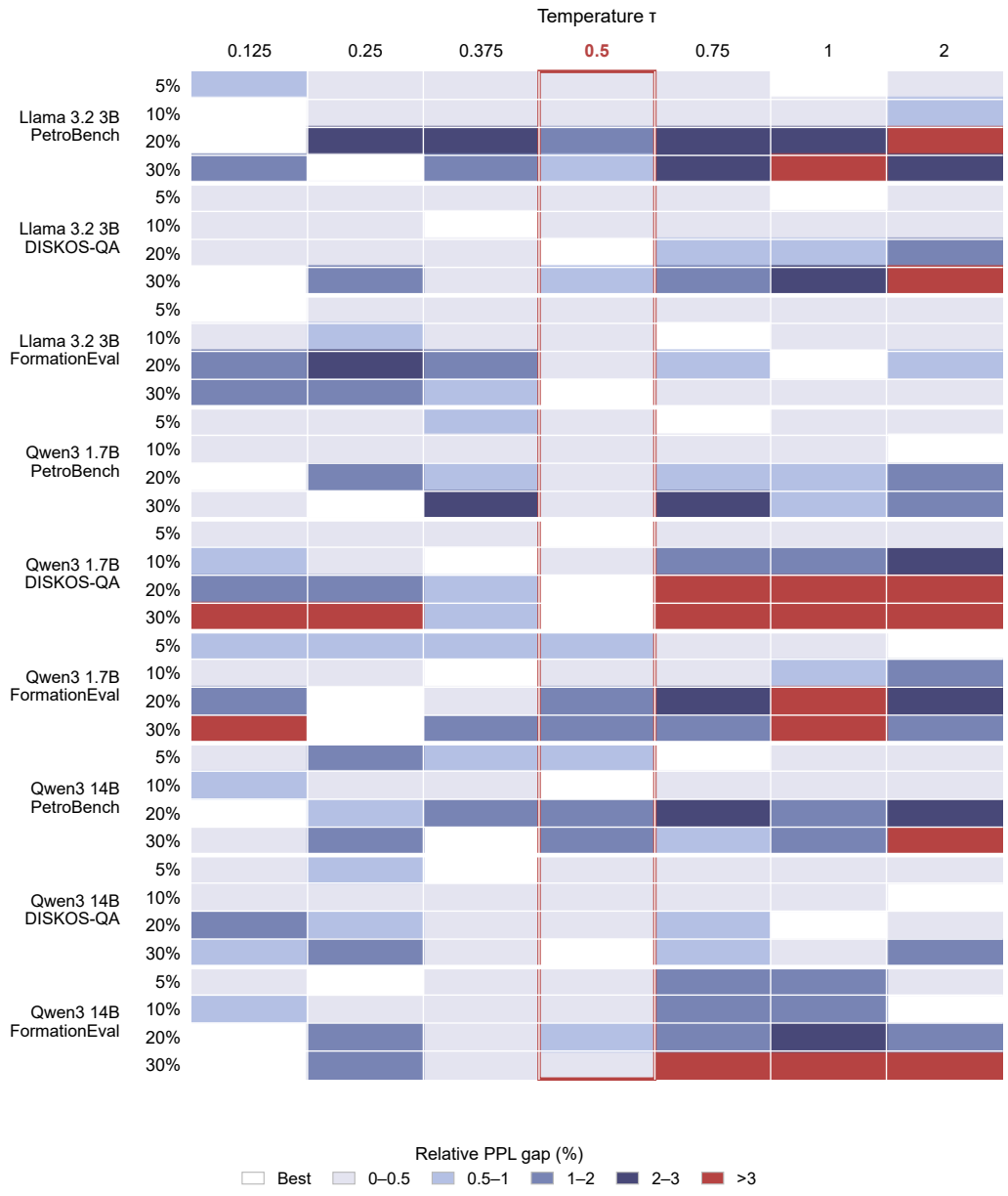

Figure 8: Relative PPL gap distributions across 36 model, dataset, and pruning ratio settings under seven candidate temperatures. White cells denote the minimum mean PPL ratio for each setting; color intensity reflects the gap relative to that row's optimal value. Red boxes mark the default temperature $\tau = 0.5$.