BOSCH: Black-Box Binary Optimization for Short-Context Attention-Head Selection in LLMs

Abbas Ghaddar* **Ivan Kobyzev*** **Boxing Chen** **Yufei Cui**
Huawei Noah's Ark Lab, Montreal Research Center, Canada
{abbas.ghaddar,ivan.kobyzev,boxing.chen,yufei.cui}@huawei.com

Abstract

Post-training hybridization of large language models (LLMs) often replaces quadratic self-attention with sliding-window attention (SWA) to reduce KV cache usage and improve latency. Existing hybridization schemes are typically defined either at the layer level (e.g., interleaving) or at the head level via static rankings from local to global. Layer-level schemes ignore that local and global dependencies are routed through heads within the same layer, while static head-level rankings suffer from *entanglement*: a head's local/global behavior can change after hybridization. We propose BOSCH, *Black-box Binary Optimization for Short-context Head Selection*, a training-free method that formulates the problem as a Large Neighborhood Search and decomposes it into three subproblems: (i) *layer-importance detection* via small-budget black-box probes, (ii) *adaptive per-layer SWA-ratio assignment* based on these sensitivities, and (iii) *grouped head-level optimization* within ratio buckets. Extensive experiments on 4 LLMs ranging from 1.7B to 30B parameters, across 4 SWA ratios, show that BOSCH consistently outperforms layer-level heuristics and 6 strong static head-level methods, with larger gains at higher SWA ratios. Under continual pretraining, BOSCH recover original long-context performance faster and to a higher level. Analysis of the selected heads reveals substantial turnover for BOSCH across different SWA ratios, underscoring the importance of performing head-level selection for each target ratio rather than relying on fixed locality rankings.

1 Introduction

Transformer self-attention (Vaswani et al., 2017) is the core of the token-mixing mechanism in state-of-the-art Large Language Models (LLMs) (Dubey et al., 2024; Team et al., 2025; Yang et al., 2025a). Due to its quadratic complexity, there has been a line of research focused on building hybrid LLMs that combine self-attention with more efficient alternatives, such as State Space Models (SSMs) (Dao and Gu, 2024; Yang et al., 2025d) or Sliding Window Attention (SWA) (Beltagy et al., 2020).

Hybridization is either performed by pretraining from scratch (OpenAI et al., 2025; Li et al., 2025b; Dong et al., 2025) or via post-training hybridization (Wang et al., 2024; Yang et al., 2025b; Gu et al., 2025). In this paper, we focus on SWA post-training hybridization of a pretrained Transformer self-attention LLM to enable efficient long-context handling. We choose SWA not only for its constant time and memory footprint, but also for its zero-shot compatibility with self-attention, which enables zero-shot hybridization and requires only minimal training for near-complete performance recovery. Given a target ratio, the main challenge lies in defining the hybridization scheme, namely which components to replace and at what granularity: layer or head level.

Previous works on LLM hybridization (Jiang et al., 2023; OpenAI, 2025; Wang et al., 2024; Yang et al., 2025b) has primarily relied on rule-based heuristics at the layer level, such as layer interleaving or begin–middle–end (BME). More recent works (Zhang et al., 2024; Gu et al., 2025) have reported improved results by deploying search-based algorithms, largely enabled by the limited search space induced by the relatively small number of layers in LLMs. However, layer-level granularity is misaligned with how transformers actually route information (Clark et al., 2019a). Local and global dependencies are handled by different attention heads within the same layer (Olsson et al., 2022; Wu et al., 2024; Tang et al., 2025). Consequently, flipping an entire layer from global to local attention can remove critical global information, leading to performance degradation.

Identifying long-context attention heads in Transformers has been an active research direc-

*Equal contribution

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 40417–40444

tion for many years, with applications in model interpretability (Clark et al., 2019b; Pascual et al., 2021), weight pruning (Kwon et al., 2022), KV-cache compression (Tang et al., 2025), attention sparsification (Wang et al., 2025), and online SWA hybridization (Donhauser et al., 2025). In the context of SWA hybridization, these static methods rank heads from local to global and then convert the most local heads according to a given ratio. However, this approach suffers from an entanglement problem: the local–global behavior of heads estimated before hybridization may change after hybridization, leading to suboptimal performance.

A straightforward solution would be to apply a search-based algorithm directly at the head level. However, this is computationally prohibitive: modern LLMs expose hundreds to low thousands of heads, making brute-force search infeasible, and black-box optimization algorithms quickly stall because each evaluation is expensive and the probability that a single random flip improves the objective drops rapidly as dimensionality grows (Shan and Wang, 2010; Frazier, 2018). Even robust black-box methods such as mesh-adaptive direct search (MADS) (Audet and Dennis, 2006) work best when each subproblem involves only tens of variables; beyond that, the number of evaluations required to find consistent improvements grows too quickly for practical budgets.

To address these issues, we formulate SWA head selection as a Large Neighborhood Search (LNS) (Shaw, 1998) problem and decompose it into 3 subproblems, making it feasible to perform the search under a realistic evaluation budget. We propose BOSCH, a black-box binary optimization search-based method that *(i)* constructs promising neighborhoods by ranking layers according to their localization sensitivity, *(ii)* assigns adaptive per-layer SWA ratios, and *(iii)* jointly optimizes the heads within each neighborhood under the target SWA ratio to produce the final SWA head-level hybridization scheme.

We conduct extensive SWA hybridization experiments on 4 models ranging from 1.7B to 30B parameters, across 4 SWA ratios, comparing against 3 layer-level and 6 head-level baseline methods. Results on 6 *needle-in-a-haystack* (NIAH) and 30 long-context QA tasks from the LongBench (Bai et al., 2024) benchmark show that our BOSCH systematically outperforms both layer-level and static head-level prior methods, with the performance gap becoming more significant at higher SWA ratios. In addition, we show that BOSCH recovers performance after continual training both faster and to a higher level than best existing methods. Our analysis reveals a correlation between method performance and the pairwise similarity of the heads selected by these methods. Moreover, we observe significant sets of heads that appear in the BOSCH SWA head set at smaller ratios (e.g., 0.5) but not at higher ratios, indicating that mitigating entanglement is crucial for performance, thereby justifying BOSCH superior performance.

2 Methodology

2.1 Problem Formalization

Let $\mathcal{M}$ be a pretrained LLM with L layers and H heads per layer. Denote the total number of heads as $N = LH$ and flatten heads to a single index $i \in \{1, \dots, N\}$. Let $z \in \{0,1\}^N$ be a binary mask, where $z_i = 1$ corresponds to the global causal self-attention and $z_i = 0$ to the sliding-window attention (SWA). Finally, denote $\mathcal{L} : (\mathcal{M}, z, \mathcal{D}) \to \mathbb{R}$ to be a loss evaluated on a calibration set $\mathcal{D}$. We formulate the SWA head selection as the following constrained binary black-box minimization problem:

$$\begin{aligned} &\min_{z \in \{0,1\}^N} \ \mathcal{L}(\mathcal{M}, z, \mathcal{D}), \\ &\text{s.t.} \quad \frac{1}{N} \sum_{i=1}^{N} \big(1 - z_i\big) = \rho, \end{aligned} \tag{1}$$

where $\rho \in [0,1]$ is the target ratio of SWA heads.

Additionally, if the model uses Grouped-Query Attention (GQA) (Ainslie et al., 2023), we enforce that all heads sharing a KV group take the same decision, i.e., we impose $z_i = z_j$ for any pair (i, j) of heads belonging to the same group. This restriction is meaningful because in GQA the keys and values are shared per group. Switching a single head to SWA while other heads in its group remain full-attention does not reduce the KV cache size; memory savings occur only when the entire KV group switches.

2.2 BOSCH

In this section, we design a loss function for the SWA hybridization problem and introduce a tailored hybridization pipeline, BOSCH, short for **B**lack-Box Binary **O**ptimization for **S**hort-**C**ontext Attention-Head **S**election.

Let $\mathcal{S}(\mathcal{M}, z, \mathcal{D})$ denote the performance score (higher the better) of the model $\mathcal{M}$ on the calibration set $\mathcal{D}$ where attention heads are localized

Layer	δ_l	w_l	r_l	g_l
L_7	-1%	0.1	0.75	0
L_6	-5%	0.4	0.25	2
L_5	-6%	0.1	0.75	0
L_4	-8%	0.2	0.5	1
L_3	-15%	0.99	0.0	–
L_2	-17%	0.2	0.5	1
L_1	-21%	0.6	0.25	2
L_0	-21%	0.0	1.0	–

Figure 1: Illustrative application of **BOSCH** to a Transformer with L=8 layers and H=8 heads per layer, targeting $\rho = 0.5\%$ SWA heads. **Left (Stage 1: layer-importance detection):** each row is a layer; blue squares are self-attention heads (z=1), red squares are SWA heads (z=0); light-red squares with a loupe indicate the layer(s) currently scored via black-box optimization, from top to bottom. **Middle (Stage 2: adaptive ratio assignment):** for each layer ℓ we compute the relative drop δ_ℓ (from the original model), normalize it to an importance weight $w_\ell \in [0, 1]$, and map it to a per-layer SWA ratio r_ℓ while respecting the global budget. **Right (Stage 3: multi-layer head selection under adaptive ratios):** layers are bucketed by ratio into groups g_ℓ and optimized jointly from more SWA-tolerant to less tolerant; the example highlights searching layers with g_ℓ=2 using r_ℓ=0.25, yielding the final head mask z. "−" indicates that either entire layer is full-attention or SWA (no optimization is applied).

according to the decision mask z. To normalize the score to be from 0 to 1 in a meaningful way, we define two anchors: $a = \mathcal{S}(\mathcal{M}, \{0\}^N, \mathcal{D})$, the performance of the total SWA model, and $b = (1+\gamma)\mathcal{S}(\mathcal{M}, \{1\}^N, \mathcal{D})$, the performance of the original full-attention model. We introduce a small factor $\gamma > 0$ to ensure a positive normalization span and reserve headroom for performance improvements beyond the original model. Then we calculate the normalized performance score as:

$$\widehat{\mathcal{S}}(\mathcal{M}, z, \mathcal{D}) \;=\; \frac{\mathcal{S}(\mathcal{M}, z, \mathcal{D}) - a}{b - a} \qquad (2)$$

leading to the design of our loss function:

$$\mathcal{L}(\mathcal{M}, z, \mathcal{D}, \rho) = -\widehat{\mathcal{S}}(\mathcal{M}, z, \mathcal{D}) \;+\; \alpha \left(\rho(z) - \rho\right)^2, \quad (3)$$

where $\rho(z) = \frac{1}{N}\sum_{i=1}^{N}(1 - z_i)$ and ρ is the target budget for the ratio of SWA heads. $\alpha > 0$ trades off validation performance against adherence to the target ratio.

Directly applying off-the-shelf black-box binary optimizers to solve problem in Equation 1 with the loss from Equation 3 is impractical for two main reasons. First, evaluating the loss is expensive: each call to $\mathcal{S}(\cdot)$ requires a LLM forward pass and typically takes several seconds, so only a small number of iterations is feasible, which severely limits its search and leads to a poor solution. Second, the search space is large. The number of binary variables is N=LH, which is in the range from a few hundred to a few thousand for billion-scale LLM. State-of-the-art black-box neighborhood methods like MADS (Audet and Dennis, 2006) explore $O(N)$ one-bit neighbors per poll, while the probability that any single bit flip yields improvement scales like $\sim 1/N$. As a result, the number of evaluations required to find improvements grows rapidly, and evaluation budgets become infeasible already beyond roughly 50 variables. In practice, toolkits like NOMAD (Audet et al., 2022) restrict themselves to subproblems with less than 50 variables, which severely constrains search efficiency and prevents effective global exploration.

Following a Large Neighborhood Search (Shaw, 1998) approach over the binary decision vector z, we split the optimization into two complementary subproblems: (i) neighborhood construction, where we identify promising subsets and per-layer budgets, and (ii) neighborhood optimization, where we jointly refine assignments within those neighborhoods. The full procedure comprises three stages, described below.

2.2.1 Layer importance detection

We first assess each layer's sensitivity to attention-head localization by iterating from the topmost layer to the bottommost layer. For each layer ℓ, we keep the current binary mask z fixed for all layers except for the current layer and run a small-budget black-box search to convert exactly $\lceil \rho H \rceil$ heads in layer ℓ to SWA, maximizing the performance

score $\mathcal{S}$. Then we include the found head configuration to update the binary mask z and proceed to the layer $\ell - 1$, so each decision is made on the updated model with previously localized upper layers. During this iteration, for every step we record the resulting best score and collect these values into $s_{\text{best}} \in \mathbb{R}^L$, which guides subsequent stages. Figure 1 (left) illustrates the search for $l = 2$ with upper layers already localized. Algorithm 1 in Appendix A gives the exact routine.

2.2.2 Adaptive ratio assignment

Given the per-layer scores s_{best} computed in Stage 1, we compute the performance drop per layer δ from the original full-attention model. Using these values we recompute the relative drop from layer to layer and rescale them to weights $w_\ell \in [0, 1]$, where lower w_ℓ means easier to localize. Layers are then sorted and bucketed into a small number of groups that map to coarse localization ratios (the target fraction of heads to convert to SWA in that layer). We reconcile these initial assignments with the global budget by shifting layers between adjacent buckets: we raise "easier" layers until the total number of localized heads matches the target, and, if needed, lower the "harder" ones. The outcome is a vector of per-layer adaptive ratios r_ℓ that allocate where localization is safest under the budget; these ratios serve as quotas for the within-layer head selection in Stage 3. See Figure 1 (middle) for an illustrative example and Algorithm 2 for detailed procedure.

2.2.3 Multi-layer head selection under adaptive ratios

Given the per-layer adaptive ratios $\{r_\ell\}$, we group layers that share the same ratio and process groups from more localizable to less (largest to smallest r). For each group, we jointly optimize the binary head decisions in its layers while keeping previously processed groups fixed. We run a small-budget black-box optimization over the concatenated heads' indices within the group to convert exactly $\lceil r_\ell H \rceil$ heads to SWA, where r_ℓ is the ratio for the current group. The resulting assignment is committed to the global mask z before proceeding to the next group. See Figure 1 (right) for an example, where we first localized layers in groups 0 and 1 with $r \geq 0.5$ and doing the search for heads in group 2 with $r = 0.25$. The exact procedure is detailed in Algorithm 3 in Appendix A.

3 Experimental Setting

In this section, we summarize the backbone models, implementation details, baseline methods from prior work, and evaluation protocols. A more detailed description of our experimental setting is provided in Appendix B.

3.1 Models

We conducted our main experiments on the Qwen3 (Yang et al., 2025a) family of models, ranging from 1.7B to 30B parameters. We selected Qwen3 because its base variants achieve state-of-the-art performance among open-weight models of same category, and have a native support for 32k sequence length.

3.2 SWA Configuration

We conduct experiments in which $\rho \in \{0.25, 0.5, 0.75, 0.875\}$ of the attention heads, groups, or layers is converted to SWA. All reported results using SWA are with a window size of 1024 (32 times smaller than the model's maximum sequence length) unless otherwise specified.

3.3 Calibration Data

Our calibration set $\mathcal{D}$ comprises synthetic *needle-in-a-haystack* (NIAH) examples (Kamradt, 2023). We compute $\mathcal{S}(\mathcal{M}, z, \mathcal{D})$ by running a prefill-only forward pass (no decoding) and, for each example, checking whether all answer tokens are predicted correctly. $\mathcal{S}(\mathcal{M}, z, \mathcal{D})$ is then the mean accuracy across examples:

$$\mathcal{S}(\mathcal{M}, z, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} \mathbf{1}[\hat{y} = y], \qquad (4)$$

where $\hat{y}$ are the model's prefill predictions for the answer tokens. We use NIAH-style data because it offers a length-controllable probe of long-range associative *recall* while minimizing train–test memorization, due to the random nature of the NIAH data-generation process.

3.4 Baselines

We compare BOSCH with baselines we constructed from prior works on long-context attention-head identification: DCAM (Clark et al., 2019a), APL (Pascual et al., 2021), FISHER (Kwon et al., 2022), QADA (Donhauser et al., 2025), RAZOR (Tang et al., 2025), and PROXY (Wang et al.,

Method	$\rho=0.25$				$\rho=0.5$				$\rho=0.75$				$\rho=0.875$			
	1.7B	**8B**	**14B**	**30B**	**1.7B**	**8B**	**14B**	**30B**	**1.7B**	**8B**	**14B**	**30B**	**1.7B**	**8B**	**14B**	**30B**
Original*	92.7	99.1	99.6	99.4	92.7	99.1	99.6	99.4	92.7	99.1	99.6	99.4	92.7	99.1	99.6	99.4
Layer-level Heuristics																
RAND	66.9	45.9	79.0	79.6	38.0	15.4	39.2	16.0	12.7	12.8	17.1	15.0	12.6	13.2	13.0	10.6
BME	40.0	30.8	41.3	47.5	11.5	12.4	11.8	18.9	11.1	12.2	11.9	12.0	11.2	12.7	12.2	10.5
INTR	74.9	72.9	41.1	69.1	19.1	19.0	14.1	12.8	12.8	12.7	11.3	22.8	12.5	12.6	11.4	11.3
Head-level Static Methods																
DCAM	80.7	95.8	96.9	92.7	12.3	84.4	87.0	81.2	10.6	22.6	30.2	41.2	10.6	13.2	13.2	9.5
APL	22.9	98.0	99.0	89.1	13.5	67.6	91.3	50.1	10.6	13.0	14.1	10.4	12.2	12.1	12.9	10.0
PROXY	50.9	97.8	97.3	88.3	23.8	71.6	68.6	76.9	11.3	31.6	42.7	29.2	11.7	13.9	13.8	10.2
QADA	83.5	95.2	83.9	80.8	60.2	75.9	70.8	61.8	38.9	46.1	19.1	34.9	16.5	13.0	17.9	10.8
RAZOR	85.5	94.1	98.9	85.0	64.8	82.4	88.0	78.4	47.8	64.9	71.3	37.2	21.9	33.9	46.9	11.6
FISHER	87.2	94.2	98.8	92.8	76.4	89.3	93.7	81.5	49.4	63.4	71.6	47.1	10.8	29.0	39.9	11.1
Ours																
BOSCH	**91.8**	**98.9**	**99.2**	**97.5**	**78.3**	**90.3**	**94.0**	**86.3**	**58.0**	**72.7**	**83.6**	**50.2**	**30.3**	**42.5**	**47.2**	**26.9**
-single	86.9	91.0	96.3	76.3	62.8	88.9	87.9	58.3	27.9	41.9	31.2	26.9	19.9	12.3	19.7	21.0
-multi	89.5	97.6	97.2	85.8	71.2	60.0	88.8	72.4	57.3	48.5	60.1	31.5	19.6	18.4	36.4	23.7
-layer	90.0	93.2	97.8	96.1	69.4	89.7	92.2	85.0	49.4	40.6	56.4	40.3	12.3	13.0	33.5	23.1

Table 1: Zero-shot average scores on the NIAH benchmark for SWA hybridization methods across 4 Qwen3 models (1.7B-30B) under 4 SWA ρ ratios. The highest and second-highest scores for each model under each ratio are highlighted in bold and underline, respectively. *For readability purposes, we repeat the original model's performance in each column (SWA hybridization is not applied to the original model).

Method	$\rho=0.25$				$\rho=0.5$				$\rho=0.75$				$\rho=0.875$			
	1.7B	**8B**	**14B**	**30B**	**1.7B**	**8B**	**14B**	**30B**	**1.7B**	**8B**	**14B**	**30B**	**1.7B**	**8B**	**14B**	**30B**
Original*	45.4	57.1	58.4	54.9	45.4	57.1	58.4	54.9	45.4	57.1	58.4	54.9	45.4	57.1	58.4	54.9
Layer-level Heuristics																
RAND	33.7	40.2	48.2	40.2	28.5	31.5	37.6	18.3	23.3	24.1	32.8	15.6	23.1	24.3	31.4	13.8
BME	31.5	31.0	32.6	28.2	21.1	24.4	15.9	16.7	20.7	28.9	29.1	12.3	20.9	26.5	30.2	14.0
INTR	36.0	43.3	43.4	42.3	28.2	34.2	25.0	25.2	19.8	29.1	30.9	14.9	19.7	23.5	30.4	13.1
Head-level Static Methods																
DCAM	32.5	43.0	52.5	40.8	22.5	33.0	42.3	31.1	18.4	26.0	33.6	21.1	17.8	24.0	31.1	10.6
APL	36.8	51.1	53.2	41.8	25.9	34.7	43.6	28.3	16.3	25.3	30.0	13.3	16.6	23.7	29.4	11.6
PROXY	34.4	50.5	49.8	38.4	24.3	33.7	39.7	27.0	19.7	29.3	32.0	17.2	18.3	27.3	29.2	12.0
QADA	36.4	40.1	43.9	35.1	28.1	30.0	36.4	27.0	25.0	21.2	31.2	20.4	21.7	23.3	28.1	12.9
RAZOR	35.1	41.2	52.2	34.5	27.9	29.2	45.6	30.5	19.8	25.7	35.9	18.8	19.8	26.7	33.0	14.3
FISHER	36.1	40.9	55.3	44.2	29.7	34.7	39.6	31.8	21.5	25.8	32.6	20.9	19.5	23.6	30.6	13.8
Ours																
BOSCH	**38.0**	**52.2**	**56.2**	**46.2**	**32.1**	**41.8**	**47.0**	**36.0**	**26.0**	**31.6**	**38.0**	**24.8**	**23.6**	**28.6**	**36.1**	**19.3**
-single	36.3	45.5	50.3	34.0	26.3	36.4	36.7	27.4	19.4	26.2	32.5	17.9	18.5	27.4	25.4	14.1
-multi	36.0	46.7	51.1	40.8	29.5	34.1	38.0	28.2	21.8	26.6	35.6	17.4	20.3	26.9	26.2	14.7
-layer	37.1	47.7	52.3	39.9	27.4	36.3	46.8	30.8	24.7	30.4	35.4	20.1	19.1	22.0	32.2	15.1

Table 2: Zero-shot average scores on the LongBench benchmark for SWA hybridization methods across 4 Qwen3 models under 4 SWA ρ ratios. Notations follow those used in Table 1.

2025). Also, we compare against the widely used *interleave* (INTR) and begin-middle-end (BME) layer-level selection heuristics (Wang et al., 2024; Yang et al., 2025c), as well as a *random* (RAND) layer selection baseline.[1] Finally, we report the results of 3 BOSCH ablations: directly using the output of stage 1 (*B-single*); using the output of stage 1 when grouping multiple layers rather than single

[1]Results are averaged over three runs.

layers (*B-multi*); and running stage 1 at layer-level granularity instead of head-level (*B-layer*), such that all layers fit in a single run.

3.5 Evaluation Benchmarks

We evaluate methods on two benchmark suites for long-context evaluation and report the average scores on each benchmark. We define the **NIAH** average as the mean performance on 6 Needle-in-a-Haystack tasks with context lengths ranging from 4k to 32k from RULER (Hsieh et al., 2024). The **LongBench** average is the unweighted average score across 6 long-context QA task categories (29 tasks) from LongBench (Bai et al., 2024), and one math reasoning task represented by GSM8K (Cobbe et al., 2021). Tables in Appendix D contain detailed per-task results for the experiments conducted in the next section.

4 Results and Analysis

4.1 Zero-Shot Results

Table 1 and Table 2 report the zero-shot performance after applying SWA hybridization with 3 layer-level heuristics, 6 head-level static methods, as well as our BOSCH and its 3 variants on the NIAH and LongBench benchmarks, respectively.

First, we notice that on both benchmarks, BOSCH consistently outperforms all baselines under every SWA ratio and model size. Among head-level static methods, FISHER and RAZOR are the strongest competitors, but they remain clearly behind BOSCH. Earlier BERT-oriented head-locality detection methods such as DCAM and APL perform substantially worse in this LLM setting, suggesting that these analysis methods do not transfer well to modern decoder-only architectures.

Second, we observe that the gap between BOSCH and the competitor methods tends to widen as the SWA ratio ρ increases, indicating that our method is particularly robust when a larger fraction of attention heads is replaced. In contrast, there is no clear and consistent similar trend with respect to model size from 1.7B to 30B parameters, pointing to model-agnostic behavior for our method. Also, we notice that commonly used layer-level heuristics for post-training hybridization, such as BME and INTR, perform poorly overall. They often fall in the same range as the RAND baseline, highlighting the limitations of simple structural heuristics.

Third, the ablations that only use the stage-one scoring (*B-single*) underperform the three-stage BOSCH pipeline by a large margin. However, expanding stage-one selection to more layers (*B-multi*) narrows the gap and approaches the performance of layer-level black-box search (*B-layer*). While these two variants still systematically fall behind BOSCH, they achieve competitive results with the best head-level static methods such as FISHER and RAZOR. This indicates the importance of mitigating entanglement effects either at the head level (*B-multi*, BOSCH) or the layer level (*B-layer*), which our black-box search framework is able to do compared to static methods.

4.2 Cross-method Heads Analysis

Figure 2 shows the Jaccard distance heatmap between SWA-selected heads for 6 head-level static methods and our BOSCH under $\rho \in \{0.5, 0.75\}$.

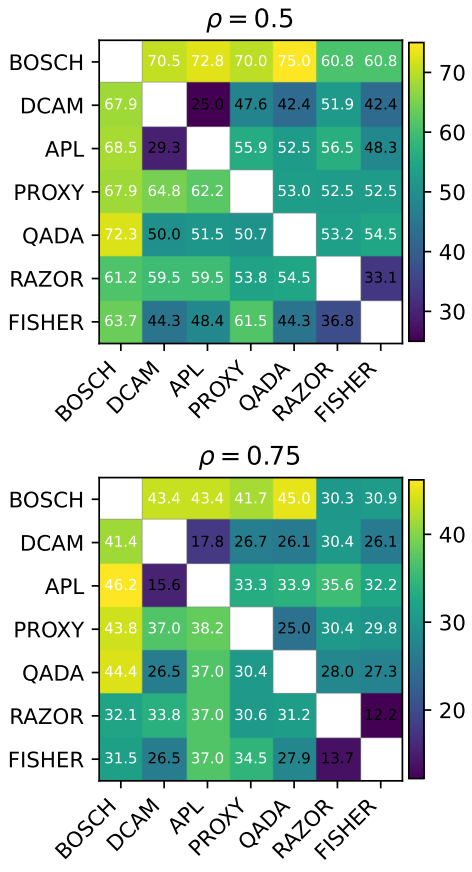

Figure 2: Jaccard distance between SWA heads selected by seven methods for $\rho = 0.25$ (upper plot) and $\rho = 0.5$ (lower plot). The distances for 8B and 14B models are shown in the lower and upper triangles, respectively.

Due to space constraints, we use the lower and upper triangles of each matrix to show results for the 8B and 14B models, respectively. Statistics for

the remaining ρ values and models are illustrated in Figure 8 in Appendix D. BOSCH exhibits relatively large Jaccard distances to all static methods, indicating that it selects a distinctly different set of SWA heads compared to these methods. Interestingly, we observe a correlation between pairwise distances and method performance: methods that achieve similar NIAH/LongBench scores tend to share more heads (i.e., have lower distances). For example, FISHER and RAZOR form a tighter pair than the other static methods (and are also the closest to BOSCH among them). In contrast, DCAM and APL are close to each other yet remain far from the high-performing group, while PROXY and QADA each select their own relatively distinct sets of heads. This may suggest that black-box, search-based methods have the flexibility to discover interactions between heads, allowing them to identify a distinct and robust set of local attention heads.

4.3 BOSCH Heads Analysis

To further validate this, we compute the fraction of heads T (the turnover ratio) that are selected by BOSCH with a small ρ_s (e.g., 0.5) but not selected by BOSCH with a larger ρ_l (e.g., 0.75). We also define the geometric turnover rate $\tilde{T}$, which normalizes T by its maximum possible value $T_{\max}(\rho_s, \rho_l)$ given only the two ρ (e.g., $T_{\max}(0.5, 0.75) = 0.5$).

Model	$0.25 \rightarrow 0.5$		$0.5 \rightarrow 0.75$		$0.75 \rightarrow 0.875$	
	T	$\tilde{T}$	T	$\tilde{T}$	T	$\tilde{T}$
1.7B	30%	30%	15%	30%	6%	36%
8B	26%	26%	15%	29%	7%	44%
14B	28%	28%	16%	31%	5%	31%
30B	29%	29%	16%	31%	6%	34%

Table 3: Raw turnover T and geometrically normalized turnover $\tilde{T}$ between BOSCH local-head sets at 2 adjacent ratios ρ. All values are reported as percentages.

Table 3 shows both statistics on 4 Qwen3 models, with 3 adjacent ρ pairs. We observe that, across all models and ρ pairs, there is a non-negligible number of heads present at ρ_s but not at ρ_l. Additionally, for the same ρ pairs, all models exhibit roughly the same range of T rates. Although T tends to decrease as ρ increases (which is expected, since $T_{\max}$ decreases), it remains consistently within the defined range of 26% (lowest) to 44% (highest), with most $\tilde{T}$ values around 30%. These trends suggest that *starting over* at each target ρ to anneal the impact of local head entanglement is crucial for performance.[2] While BOSCH must restart at each target ρ, making it more expensive than static methods that run local-head detection and ranking only once, this added cost is justified by its consistently higher performance.

4.4 Continual Training Results

Although SWA hybridization yields substantial latency reductions,[3] it also leads to siginficant performance degradation, particularly at higher ρ. We therefore conduct continual pretraining experiments with $\rho = 0.75$, which provides a favorable efficiency–quality trade-off, to evaluate how much of the original performance can be recovered after hybridization.

Figure 3 presents intermediate NIAH and LongBench scores for Qwen3-8B-Base and Qwen3-14B-Base during continual pretraining on 2.5B tokens for 4 methods: INTR and *B-layer* (layer-level), BOSCH and the strongest head-level static baseline FISHER. The training setup is detailed in § B.5, and full results, including those for 1.7 and 30B models, are reported in Appendix D.

On the one hand, we observe that head-level methods can substantially recover the original model performance better than layer-level methods. Both BOSCH and FISHER nearly close the NIAH gap and significantly reduce the LongBench gap after 2.5B additional tokens, whereas layer-level methods remain clearly underperforming. However, our black-box layer-level variant *B-layer* consistently outperforms the INTR heuristic, suggesting that black-box search is beneficial even when restricted to layer granularity.

On the other hand, BOSCH retains a clear advantage over FISHER after continual pretraining, especially on LongBench, indicating that better head masks at initialization translate into better downstream generalization even when extra training is allowed. Interstingly, even with with a high $\rho = 0.75$, 2.5B tokens are sufficient to almost fully recover long-context recall on NIAH and to substantially narrow the gap on LongBench. This is encouraging when compared to architecture-level hybridization approaches such as MambaInLlama or Jet-Nemotron, which require orders of magnitude more training tokens and pipeline training. Our results suggest that post-hoc SWA hybridization, combined with a modest amount of continual

[2] We further confirm this finding with additional experiments in § C.1.

[3] See § C.4 for a latency and memory analysis.

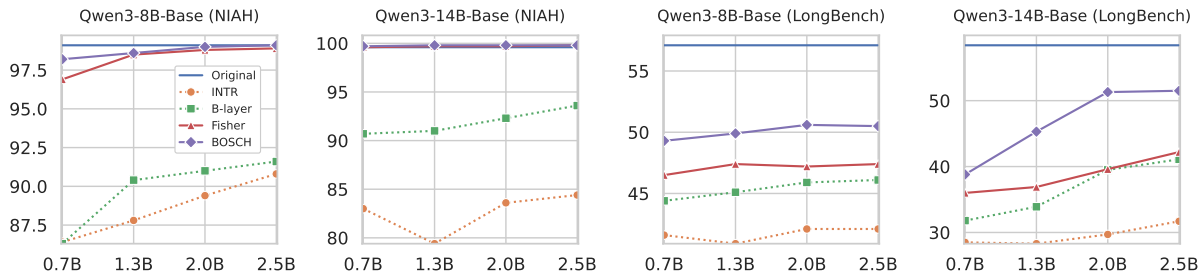

Figure 3: NIAH (first 2 plots) and LongBench (last 2 plots) performances (y-axis) for Qwen3-8B-Base and Qwen3-14B-Base models as a function of intermediate checkpoints during continual pretraining on 2.5B tokens (x-axis). We report scores for two layer-level and two head-level SWA hybridization methods using $\rho = 0.75$. The performance of the full-attention original models (without additional training) is shown as a straight line with no markers.

pretraining, can be an efficient and practical alternative for deploying long-context–efficient variants of existing LLMs.

5 Related Work

There has been a surge of hybrid models that combine Transformer layers with more efficient modules like linear SSM/RNN blocks (Dao and Gu, 2024; Lieber et al., 2024; Dong et al., 2024; Li et al., 2025a; Qwen Team, 2025) or sliding-window attention (SWA) (OpenAI, 2025; Jiang et al., 2023). These models are trained from scratch and typically require substantial pretraining compute. To reduce compute, continual-training hybridization have been proposed (Wang et al., 2024; Yang et al., 2025b; Lu et al., 2025): starting from a pre-trained transformer it replaces full-attention with more-efficient blocks post-hoc. In these works, the placement decision is at the layer level via a fixed interleaving schedule.

Recently, Jet-Nemotron (Gu et al., 2025) approaches hybridization via neural architecture search: starting from a pre-trained full-attention Transformer, the authors freeze MLP weights, search over placements of residual full-attention layers and candidate linear-attention blocks, and use beam search on a calibration set to select configurations for a target hybridization ratio. Our method can replace beam search with an advanced black-box search algorithm and apply a head-level linear-attention hybridization warm-up instead of layer-level mixing.

A complementary line of work targets training-free head-level KV-cache reduction. RazorAttention (Tang et al., 2025) is a static method that identifies attention heads important for long-context processing (so called “retrieval heads") using a synthetic repeated-token probe and truncates distant tokens for the rest. Donhauser et al. (2025) propose a dynamic per-token detector that classify long context-relevant heads based on the attention mass condensation in the window near the boundary. Unlike these detection-based techniques, BOSCH directly optimizes a binary head mask under explicit budgets, taking into account the inter-head dependency and the entangled behavior.

Finally, post-training pruning removes attention heads without retraining: Kwon et al. (2022) propose Fisher-guided structured pruning with mask rearrangement and per-layer output reconstruction, and ProxyAttn (Wang et al., 2025) estimates block importance training-free by pooling scores from representative heads and allocating a dynamic sparsity budget. In contrast to these per-component, statistics-driven methods, our approach formulates selection as black-box binary optimization enabling joint optimization across heads.

6 Conclusion

In this work, we propose BOSCH, a training-free black-box binary optimization framework for short-context attention-head selection for LLMs SWA post-training hybridization. We decompose the problem into 3 stages: layer-importance detection, adaptive per-layer SWA-ratio assignment, and grouped head-level search. BOSCH consistently outperform layer-level heuristics and strong static head-level baselines on long-context benchmarks across multiple models sizes and SWA ratios. As future work, we plan to extend BOSCH to other hybrid primitives incorporating SSM layers, where a key challenge is the current inability to perform zero-shot search. We also aim to adapt our framework to support Multi-Latent Attention (Liu et al., 2024), in which all attention heads are compressed into a single continuous vector.

Limitations

Our experiments are conducted exclusively on the Qwen3 family in order to prioritize the systematicity of our evaluation. For instance, other model families (Team et al., 2024; Touvron et al., 2023; Dubey et al., 2024; Abdin et al., 2024) either natively support relatively shorter maximum sequence lengths (e.g., 4k or 8k tokens), do not offer multiple model sizes, or are only available as instruction-tuned variants. Due to computational constraints, we limit our exploration of continual pretraining to a single data mixture and at most 2.5B additional tokens, and we do not experiment with models larger than 30B parameters. Finally, we do not study how SWA hybridization interacts with other efficiency techniques such as quantization, KV-cache compression, or weight pruning, leaving their compatibility and potential compounding effects for future work.

Acknowledgments

We thank the anonymous reviewers for their insightful comments.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

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

Charles Audet and J. E. Dennis. 2006. Mesh adaptive direct search algorithms for constrained optimization. *SIAM Journal on Optimization*, 17(1):188–217.

Charles Audet, Sébastien Le Digabel, Viviane Rochon Montplaisir, and Christophe Tribes. 2022. Algorithm 1027: NOMAD Version 4: Nonlinear Optimization with the MADS Algorithm. *ACM Transactions on Mathematical Software*, 48(3):35:1–35:22.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137, Bangkok, Thailand. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *ArXiv*, abs/2004.05150.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, and 1 others. 2020. Piqa: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI conference on artificial intelligence*, 34(05):7432–7439.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019a. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019b. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv e-prints*, pages arXiv–2110.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations*.

Tri Dao and Albert Gu. 2024. Transformers are ssms: generalized models and efficient algorithms through structured state space duality. In *Proceedings of the 41st International Conference on Machine Learning*, pages 10041–10071.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Xin Dong, Yonggan Fu, Shizhe Diao, Wonmin Byeon, Zijia Chen, Ameya Sunil Mahabaleshwarkar, Shih-Yang Liu, Matthijs Van Keirsbilck, Min-Hung Chen, Yoshi Suhara, Yingyan Celine Lin, Jan Kautz, and Pavlo Molchanov. 2025. Hymba: A hybrid-head architecture for small language models. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

Xin Dong, Yonggan Fu, Shizhe Diao, Wonmin Byeon, Zijia Chen, Ameya Sunil Mahabaleshwarkar, Shih-Yang Liu, Matthijs Van Keirsbilck, Min-Hung Chen, Yoshi Suhara, and 1 others. 2024. Hymba: A hybrid-head architecture for small language models. *arXiv preprint arXiv:2411.13676*.

Konstantin Donhauser, Charles Arnal, Mohammad Pezeshki, Vivien Cabannes, David Lopez-Paz, and Kartik Ahuja. 2025. Unveiling simplicities of attention: Adaptive long-context head identification. *ArXiv*, abs/2502.09647.

Stefan Droste, Thomas Jansen, and Ingo Wegener. 2002. On the analysis of the (1+ 1) evolutionary algorithm. *Theoretical Computer Science*, 276(1-2):51–81.

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

Peter I. Frazier. 2018. A tutorial on bayesian optimization. *ArXiv*, abs/1807.02811.

Tianyu Gao, Alexander Wettig, Howard Yen, and Danqi Chen. 2024. How to train long-context language models (effectively). *arXiv preprint arXiv:2410.02660*.

Yuxian Gu, Qinghao Hu, Haocheng Xi, Junyu Chen, Shang Yang, Song Han, and Han Cai. 2025. Jet-nemotron: Efficient language model with post neural architecture search. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. Ruler: What's the real context size of your long-context language models? In *Proceedings of the First Conference on Language Modeling (COLM)*, COLM, Philadelphia, PA, USA.

Pin-Lun Hsu, Yun Dai, Vignesh Kothapalli, Qingquan Song, Shao Tang, Siyu Zhu, Steven Shimizu, Shivam Sahni, Haowen Ning, and Yanning Chen. 2024. Liger kernel: Efficient triton kernels for llm training. *arXiv preprint arXiv:2410.10989*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Gregory Kamradt. 2023. Needle In A Haystack - pressure testing LLMs. *Github*.

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

Sébastien Le Digabel. 2011. Algorithm 909: Nomad: Nonlinear optimization with the mads algorithm. *ACM Transactions on Mathematical Software*, 37(4):44:1–44:15.

Aonian Li, Bangwei Gong, Bo Yang, Boji Shan, Chang Liu, Cheng Zhu, Chunhao Zhang, Congchao Guo, Da Chen, Dong Li, Enwei Jiao, Gengxin Li, Guojun Zhang, Haohai Sun, Houze Dong, Jiadai Zhu, Jiaqi Zhuang, Jiayuan Song, Jin Zhu, and 70 others. 2025a. Minimax-01: Scaling foundation models with lightning attention. *ArXiv*, abs/2501.08313.

Aonian Li, Bangwei Gong, Bo Yang, Boji Shan, Chang Liu, Cheng Zhu, Chunhao Zhang, Congchao Guo, Da Chen, Dong Li, and 1 others. 2025b. Minimax-01: Scaling foundation models with lightning attention. *arXiv preprint arXiv:2501.08313*.

Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Haim Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, Omri Abend, Raz Alon, Tomer Asida, Amir Bergman, Roman Glozman, Michael Gokhman, Avshalom Manevich, Nir Ratner, Noam Rozen, and 3 others. 2024. Jamba: A hybrid transformer-mamba language model. *ArXiv*, abs/2403.19887.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Peng Lu, Ivan Kobyzev, Mehdi Rezagholizadeh, Boxing Chen, and Philippe Langlais. 2025. Regla: Refining gated linear attention. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2884–2898.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2018. Mixed precision training. In *In International Conference on Learning Representations*.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova Dassarma, T. J. Henighan, Benjamin Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, John Kernion, Liane Lovitt, and 7 others. 2022.

In-context learning and induction heads. *ArXiv*, abs/2209.11895.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *ArXiv*, abs/2508.10925.

OpenAI and 1 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Damian Pascual, Gino Brunner, and Roger Wattenhofer. 2021. Telling BERT's full story: from local attention to global aggregation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 105–124, Online. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2024. Yarn: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations*.

Qwen Team. 2025. Qwen3-next: Towards ultimate training & inference efficiency. Technical Report.

Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. 2024. Flashattention-3: Fast and accurate attention with asynchrony and low-precision. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Shih Shan and Guo Guang Wang. 2010. Survey of modeling and optimization strategies to solve high-dimensional design problems with computationally-expensive black-box functions. *Structural and Multidisciplinary Optimization*, 41(2):219–241.

P. Shaw. 1998. Using constraint programming and local search methods to solve vehicle routing problems. In *CP-98 (Fourth International Conference on Principles and Practice of Constraint Programming)*, volume 1520 of *Lecture Notes in Computer Science*, pages 417–431. Springer.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Hanlin Tang, Yang Lin, Jing Lin, Qingsen Han, Danning Ke, Shikuan Hong, Yiwu Yao, and Gongyi Wang. 2025. Razorattention: Efficient kv cache compression through retrieval heads. In *The Thirteenth International Conference on Learning Representations*.

Yehui Tang, Kai Han, Fangcheng Liu, Yunsheng Ni, Yuchuan Tian, Zheyuan Bai, Yi-Qi Hu, Sichao Liu, SHANGLING JUI, and Yunhe Wang. 2024. Rethinking optimization and architecture for tiny language models. In *Forty-first International Conference on Machine Learning*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, and 1 others. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Junxiong Wang, Daniele Paliotta, Avner May, Alexander M. Rush, and Tri Dao. 2024. The mamba in the llama: Distilling and accelerating hybrid models. In *Advances in Neural Information Processing Systems*, volume 37, pages 62432–62457. Curran Associates, Inc.

Yixuan Wang, Huang He, Siqi Bao, Hua Wu, Haifeng Wang, Qingfu Zhu, and Wanxiang Che. 2025. Proxy-attn: Guided sparse attention via representative heads. *ArXiv*, abs/2509.24745.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. 2024. Retrieval head mechanistically explains long-context factuality. *ArXiv*, abs/2404.15574.

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, and 41 others. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Mingyu Yang, Mehdi Rezagholizadeh, Guihong Li, Vikram V. Appia, and Emad Barsoum. 2025b. Zebra-llama: Towards extremely efficient hybrid models. In *Advances in Neural Information Processing Systems*, volume 39.

Mingyu Yang, Mehdi Rezagholizadeh, Guihong Li, Vikram V. Appia, and Emad Barsoum. 2025c. Zebra-llama: Towards extremely efficient hybrid models. *ArXiv*, abs/2505.17272.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. 2025d. Gated delta networks: Improving mamba2 with delta rule. In *The Thirteenth International Conference on Learning Representations*.

Jun Zhang, Jue Wang, Huan Li, Lidan Shou, Ke Chen, Gang Chen, and Sharad Mehrotra. 2024. Draft & verify: Lossless large language model acceleration via self-speculative decoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11263–11282, Bangkok, Thailand. Association for Computational Linguistics.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, and 1 others. 2023. Pytorch fsdp: Experiences on scaling fully sharded data parallel. *Proceedings of the VLDB Endowment*, 16(12):3848–3860.

A BOSCH Algorithms

We present algorithms for Stages 1–3 introduced in § 2. To this end, we slightly generalize the loss in Formula (3) to accommodate the smaller-cardinality subproblems arising in the neighborhood optimization steps. The general form of the loss is unchanged but we modify the penalty term. Let N be the total number of heads and $\mathfrak{I} \subset \{1, \ldots, N\}$ is a subset of heads. Define the loss as:

$$\begin{aligned} \mathcal{L}(\mathcal{M}, z, \mathcal{D}, \bar{\rho}, \mathfrak{I}) &= -\widehat{\mathcal{S}}(\mathcal{M}, z, \mathcal{D}) \; + \; \alpha\big(\bar{\rho}_{\mathfrak{I}}(z) - \bar{\rho}\big)^2, \\ \text{where} \quad \bar{\rho}_{\mathfrak{I}}(z) &= \frac{1}{|\mathfrak{I}|} \sum_{i \in \mathfrak{I}} (1 - z_i). \end{aligned} \tag{5}$$

As for the original loss, $\widehat{\mathcal{S}}$ is the normalized score function from Formula 2, $\bar{\rho}$ is the target ratio of SWA heads for the subproblem we are solving, and $\alpha > 0$ trades the validation performance against the adherence to the target ratio.

For Stage 1 we optimize at the layer granularity. For each layer ℓ the index set of heads is:

$$\mathfrak{I}_\ell = \{(\ell - 1) \cdot H + 1, \ldots, (\ell - 1) \cdot H + H\}. \tag{6}$$

Algorithm 1 Stage 1

Require: Model $\mathcal{M}$, calibration set $\mathcal{D}$, target ratio of SWA heads ρ
Require: Layers L, heads H, Scoring Function $\mathcal{S}(\cdot)$, Black-Box Optimizer BBO
$N \leftarrow LH$; $\quad z \leftarrow \mathbf{1} \in \{0,1\}^N$; $\quad s_{\text{best}} \leftarrow \mathbf{0} \in \mathbb{R}^L$
for $\ell = L,\ L{-}1,\ \ldots,\ 1$ **do**
  $i_0 \leftarrow (\ell - 1) \cdot H + 1$; $\quad i_1 \leftarrow i_0 + H$; $\quad \mathfrak{I}_\ell = \{i_0, \ldots, i_1\}$ ▷ indices of heads in the layer ℓ
  $\mathcal{U} \leftarrow \{\, u \in \{0,1\}^N : \; u[k] = z[k] \quad \forall k \notin \mathfrak{I}_\ell, \}$ ▷ fix the mask outside of the layer ℓ
  $z \leftarrow \min_{u \in \mathcal{U}} \mathcal{L}(\mathcal{M}, u, \mathcal{D}, \rho, \mathfrak{I}_\ell)$ ▷ optimize the loss from Eq. 5 with the BBO
  $s_{\text{best}}[\ell] \leftarrow \mathcal{S}(\mathcal{M}, z, \mathcal{D})$
return s_{best}

For Stage 3 we optimize for the heads based on the adaptive ratios. We explicitly form the index set by unifying the groups with the same ratio and use this ratio as the target budget for SWA heads.

Algorithm 3 Stage 3

Require: Model $\mathcal{M}$, calibration set $\mathcal{D}$
Require: Layers L, heads H, per-layer adaptive ratios $\{r_\ell\}_{\ell=1}^{L}$
Require: Scoring Function $\mathcal{S}(\cdot)$, Black-Box Optimizer BBO
$N \leftarrow LH$; $\quad z \leftarrow \mathbf{1} \in \{0,1\}^N$
$M_\ell \leftarrow \lceil r_\ell H \rceil$ for $\ell = 1, \ldots, L$
Sort layers by r_ℓ in descending order and split into groups $G_1, \ldots, G_K$ with identical r_ℓ.
for $g = 1$ to K **do**
  $\mathfrak{I}_g \leftarrow \bigcup_{\ell \in G_g} \{\ell H + 1,\ \ell H{+}2,\ \ldots,\ \ell H{+}H\}$ ▷ indices of all heads in the current group
  $\mathcal{U} \leftarrow \Big\{ u \in \{0,1\}^N : \; u[k] = z[k] \quad \forall k \notin \mathfrak{I}_g \}$ ▷ fix the mask outside of the current group
  $z \leftarrow \min_{u \in \mathcal{U}} \mathcal{L}(\mathcal{M}, u, \mathcal{D}, r_l, \mathfrak{I}_g)$ ▷ optimize the loss from Eq. 5 with the BBO
return z

Algorithm 2 Stage 2

Require: Layer scores $s_{\text{best}} \in \mathbb{R}^L$, score of the full-attention model s_{orig}
Require: Target ratio of SWA heads ρ, layers L, heads per layer H
Require: Buckets B, partition of the unit interval $\{b_0 < \cdots < b_{B-1}\} \subset [0, 1]$
Require: Percentiles p_{low}=2.5, p_{high}=97.5

$N \leftarrow LH$; $M_{\text{target}} \leftarrow \rho N$
$\delta_\ell \leftarrow \dfrac{s_{\text{best}}[\ell] - s_{\text{orig}}}{s_{\text{orig}}}$ for ℓ=1:L; $\delta_{L+1} \leftarrow 0$ ▷ compute absolute performance drop
$d_\ell \leftarrow \delta_\ell - \delta_{\ell+1}$ for ℓ=1:L ▷ compute per layer relative performance drop
$q_{\text{lo}}, q_{\text{hi}} \leftarrow p_{\text{low}}, p_{\text{high}}$ percentiles of $\{d_\ell\}$
for $\ell = 0$ to $L-1$ **do**
    $\tilde{d}_\ell \leftarrow \min\{\max(d_\ell, q_{\text{lo}}), q_{\text{hi}}\}$ ▷ clip to central percentiles
    $w_\ell \leftarrow \dfrac{\tilde{d}_\ell - q_{\text{lo}}}{q_{\text{hi}} - q_{\text{lo}}}$ ▷ compute normalized weights per layer
Sort layers by w_ℓ in descending order and split into B equal groups $G_0, \ldots, G_{B-1}$
Set $\mathsf{rank}[\ell] \leftarrow j$ for $\ell \in G_j$
$M_{\text{now}} \leftarrow \sum_\ell b_{\mathsf{rank}[\ell]} H$
$\Delta \leftarrow M_{\text{target}} - M_{\text{now}}$ ▷ head count gap; positive means more head can be localized
$\mathsf{easy} \leftarrow$ layers sorted by w_ℓ ascending; $\mathsf{hard} \leftarrow$ layers sorted by w_ℓ descending
while $\Delta \neq 0$ **do**
    if $\Delta > 0$ **then**
        for ℓ in easy **do**
            if $\mathsf{rank}[\ell] < B-1$ **then**
                $r \leftarrow \mathsf{rank}[\ell]$; $\delta \leftarrow H\,(b_{r+1} - b_r)$
                if $\delta \leq \Delta$ **then**
                    $\mathsf{rank}[\ell] \leftarrow r + 1$; $\Delta \leftarrow \Delta - \delta$
                    if $\Delta = 0$ **then**
                        break
    else
        for ℓ in hard **do**
            if $\mathsf{rank}[\ell] > 0$ **then**
                $r \leftarrow \mathsf{rank}[\ell]$; $\delta \leftarrow H\,(b_r - b_{r-1})$
                if $\delta \leq -\Delta$ **then**
                    $\mathsf{rank}[\ell] \leftarrow r - 1$; $\Delta \leftarrow \Delta + \delta$
                    if $\Delta = 0$ **then**
                        break
$r_\ell \leftarrow b_{\mathsf{rank}[\ell]}$ for all ℓ
return $\{r_\ell\}_{\ell=1}^{L}$ ▷ adaptive ratios

B Experimental Setting

B.1 Models

Model	#P	#L	#QH	#KVH	ML
Q3-1.7B	1.7	28	16	8	32k
Q3-8B	8.2	36	32	8	32k
Q3-14B	14.8	40	40	8	32k
Q3-30B-A3B	30.5	48	32	4	40k

Table 4: Qwen3-Base backbones used in our experiments. We report parameter count (#P, in billions), number of layers (#L), number of query attention heads (#QH), number of key/value heads (#KVH), and maximum length (ML, tokens).

Table 4 summarizes the main characteristics of the Qwen3-Base models used in this study. All models use grouped-query attention (GQA) for efficient self-attention computation and natively support sequence lengths of at least 32k tokens (the 30B models support up to 40k tokens). The largest 30B model is a Mixture-of-Experts (MoE) model (Shazeer et al., 2017) with only 3B parameters activated during inference, while all other models are dense. Therefore, it sometimes underperforms fully dense models (e.g., the 14B) in some settings.

B.2 BOSCH Implementation Details

We use the MADS algorithm (Audet and Dennis, 2006), as implemented in the NOMAD toolkit (Le Digabel, 2011; Audet et al., 2022), as our binary black-box optimizer. Its surrogate quadratic-model search is only enabled for problems with at most 50 variables, which otherwise backoff to OnePlusOne (Droste et al., 2002) random variable selection. For example, with 50 variables we can group at most four layers (i.e., enforce $\max|G_g| \leq 4$ in Algorithm 3) for the 1.7B, 8B, and 14B Qwen3 models, each with 8 KV heads. If the number of layers that number, G_g of Algorithm 3 is partitioned into smaller equally size subsets.

Let κ denote the per-layer iteration budget, we we use $\kappa = 100$ in all experiments unless noted. In all experiments with BOSCH, we allocate κ iterations per layer across both BOSCH stages and in the layer-level ablation (*B-layer*). For instance, line 5 in Algorithm 1 runs for κ iterations each time it is called; line 7 in Algorithm 3 runs for $\kappa\,|G_g|$ iterations for a group G_g; and the layer-level ablation runs for $\kappa\,L$ iterations. If, in any case, the number of possible candidates is less than or equal to the total iteration budget, we fall back to a brute-force algorithm that evaluates all candidates. Finally, we set α and γ in § 2.2 to 100 and 0.2, respectively.

BOSCH has two implementation modes: search and deploy. The search mode is used during optimization on the calibration set $\mathcal{D}$; for each forward pass and candidate mask z, hybridization is applied only at the locations indicated by z. The deploy mode is used after the final mask z has been selected, for both inference and training. Both modes are implemented in PyTorch (Paszke et al., 2019) on top of the Transformers library (Wolf et al., 2020) and are fully compatible with FlashAttention-2 (Dao, 2023) for efficient computation of both self-attention and SWA.

Search mode We load the original self-attention model and, for each pass over $\mathcal{D}$ with a given mask z, handle any layer ℓ that requires hybrid SA/SWA heads as follows. After computing the QKV projections, we partition the heads into SA and SWA subsets according to z. We then run the SA and SWA kernels in parallel (on two streams) and concatenate their outputs along the head dimension to form the attention output tensor. Before applying the output projection, we permute the rows of W_o to match the $[\,\text{SA},\ \text{SWA}\,]$ head ordering of the concatenated tensor, and then apply W_o. This procedure requires loading the model once and allows z to be updated efficiently after each full pass over $\mathcal{D}$. When grouped-query attention (GQA) is used, slicing respects KV-group boundaries: all heads in a KV group share the same decision.

Deploy mode Given the final mask z, we materialize the per-layer hybrid modules once at model initialization. For each hybrid layer ℓ, we slice the QKV projection weight matrices along the head dimension into self-attention (SA) and sliding-window attention (SWA) subsets according to z, and permute the rows of W_o to match the $[\,\text{SA},\ \text{SWA}\,]$ head ordering. At inference or training time, we compute the SA and SWA attention outputs in parallel (e.g., via separate kernel launches), concatenate them along the head dimension, and apply the output projection with W_o. When using GQA, slicing respects KV-group boundaries so that all heads within a KV group share the same decision.

B.3 Baselines

We define 6 training-free long-context head-selection methods built on prior works. These baselines were not originally designed for selecting SWA heads for model hybridization, but rather for other related tasks such as model interpretability (Clark et al., 2019b; Pascual et al., 2021), weight pruning (Kwon et al., 2022), KV-cache compression (Tang et al., 2025), attention sparsification (Wang et al., 2025), and online SWA hybridization (Donhauser et al., 2025). Let denote by $\alpha_{\ell,h}(t,j) \in [0,1]$ the attention probability at layer $\ell \in \{1,\dots,L\}$ and head $h \in \{1,\dots,H\}$ from source position j to target position t (causal, $j \le t$). Lags are denoted by $d = t - j$, where all lag are self-included (d=0). With GQA, we compute each KV-group's score as the mean over the heads in that group, as it consistently outperformed max pooling across methods in preliminary experiments. Each baseline yields a per-head score $s_{\ell,h}$ where larger values indicate more local unless noted. Given a target SWA ratio ρ, we rank heads (or groups) by from local to global and apply SWA to the top $\lceil \rho N \rceil$ (setting $z_i = 0$), leaving the rest as full self-attention ($z_i = 1$). SWA window W is set to the same value (e.g. 1024) used when experimenting with our BOSCH method.

B.3.1 Distance-Conditioned Attention Mass (DCAM)

Clark et al. (2019b) studied BERT (Devlin et al., 2019) attention head specialization patterns by conditioning on relative token distance by quantifying how much mass fell on nearby versus distant tokens. We adapt their method to obtain a single locality score per head ($s_{\ell,h}$) by aggregating attention scores by lag and normalizing to a per-head histogram:

$$M_{\ell,h}(d) \;=\; \sum_{(x,t)\in\mathcal{D}} \alpha_{\ell,h}(t,\, t{-}d), \qquad p_{\ell,h}(d) \;=\; \frac{M_{\ell,h}(d)}{\sum_{u\ge 0} M_{\ell,h}(u)}, \tag{7}$$

Given an SWA window W, we define the DCAM locality score as the cumulative mass inside the window:

$$s_{\ell,h} \;=\; \sum_{d=0}^{W-1} p_{\ell,h}(d), \tag{8}$$

where higher $s_{\ell,h}$ indicates a more local head. We compute $s_{\ell,h}$ on the calibration set $\mathcal{D}$ and rank heads in descending order.

B.3.2 Answer-Centric Peak Lag (APL)

Pascual et al. (2021) analyze heads via distance-conditioned attention patterns to characterize local vs. global heads in BERT-like models. We adapt this idea to our NIAH calibration set $\mathcal{D}$ by anchoring on the last answer token and measuring how far each head's peak attention lies from it.

For each example $(x,y) \in \mathcal{D}$, let $t^\star(x)$ be the last answer token and define the peak lag

$$j^\star_{\ell,h}(x) = \arg\max_{j \le t^\star(x)} \alpha_{\ell,h}\big(t^\star(x), j\big), \qquad \lambda_{\ell,h}(x) = t^\star(x) - j^\star_{\ell,h}(x). \tag{9}$$

We aggregate across examples and normalize by the SWA window W to obtain a locality score (larger $s_{\ell,h}$ means more local head):

$$\bar{\lambda}_{\ell,h} = \mathrm{median}_{x\in\mathcal{D}}\, \lambda_{\ell,h}(x), \qquad s_{\ell,h} = 1 - \min\left(1, \frac{\bar{\lambda}_{\ell,h}}{W}\right). \tag{10}$$

B.3.3 Proxy Attention (PROXY)

PROXYATTN (Wang et al., 2025) is a training-free guided sparse-attention algorithm that pools a few representative *proxy* heads to compute unified token/block importance scores and assigns head-specific dynamic sparsity budgets. We adapt the their budget estimator into a per-head locality score on $\mathcal{D}$. For each example $(x,y) \in \mathcal{D}$ and *target position* t, partition the *source positions* $j \in \{1,\dots,t\}$ into n_B

contiguous blocks of size B. Form a pooled, layer-wise proxy attention by taking the maximum over a small proxy set $\mathcal{P}_\ell \subseteq \{1, \ldots, H\}$ (default $|\mathcal{P}_\ell|$=1):

$$\tilde{\alpha}_\ell(t, j) \;=\; \max_{h \in \mathcal{P}_\ell} \alpha_{\ell,h}(t, j), \tag{11}$$

then we compute pooled block masses and a common ranking:

$$S_\ell(b, t) \;=\; \sum_{j \in \text{block } b} \tilde{\alpha}_\ell(t, j), \qquad \text{then sort } \{S_\ell(b, t)\}_{b=1}^{n_B} \text{ in descending order.} \tag{12}$$

For each head h, let $m_{\ell,h}(t)$ be the smallest number of top-ranked blocks whose own cumulative mass reaches a target fraction $\gamma \in (0, 1]$:

$$m_{\ell,h}(t) \;=\; \min\Big\{m' \colon \sum_{b=1}^{m'} \sum_{j \in \text{block } b} \alpha_{\ell,h}(t, j) \;\geq\; \gamma \sum_{b=1}^{n_B} \sum_{j \in \text{block } b} \alpha_{\ell,h}(t, j)\Big\}. \tag{13}$$

We define the per-row fractional budget $q_{\ell,h}(t) = \frac{m_{\ell,h}(t)}{n_B}$ and the head score:

$$s_{\ell,h} \;=\; \mathbb{E}_{(x,t)\in\mathcal{D}}\big[q_{\ell,h}(t)\big], \tag{14}$$

where larger $s_{\ell,h}$ means that the mass is concentrated in fewer nearby blocks and consequently is more local head. Therefore, we rank heads by ascending order.

B.3.4 Query-ADaptive Attention Criterion (QADA)

QADA (Donhauser et al., 2025) aims to estimate, for each token and head, how much attention mass lies inside a local window without explicitly computing full-context scores. They do so by modeling the non-local (global) token scores with a second-moment Gaussian approximation. We adapt QADA to produce a single locality score per head only on our calibration set $\mathcal{D}$. For a given example $(x, y) \in \mathcal{D}$, layer ℓ, head h, and target position t, let $q_{\ell,h}(t) \in \mathbb{R}^{d_k}$ and $k_{\ell,h}(j) \in \mathbb{R}^{d_k}$ be the query and key vectors. We define the *local* window and its complement:

$$\mathcal{N}_W(t) = \{\, j :\; \max(1, t{-}W{+}1) \leq j \leq t \,\}, \qquad \mathcal{F}_W(t) = \{\, j :\; 1 \leq j \leq t{-}W \,\}. \tag{15}$$

then we compute the exact unnormalized local softmax mass:

$$A_{\text{local}}(t) \;=\; \sum_{j \in \mathcal{N}_W(t)} \exp\big(\tau\, q_{\ell,h}(t)^\top k_{\ell,h}(j)\big). \tag{16}$$

To approximate the *non-local* mass, we model keys in $\mathcal{F}_W(t)$ as Gaussian along the query direction using a small boundary buffer $\mathcal{B}_B(t) = \{t{-}W{-}B{+}1, \ldots, t{-}W\} \subseteq \mathcal{F}_W(t)$ to estimate second moments:

$$\hat{\mu}_{\ell,h}(t) = \frac{1}{|\mathcal{B}_B(t)|} \sum_{j \in \mathcal{B}_B(t)} k_{\ell,h}(j), \qquad \hat{\Sigma}_{\ell,h}(t) = \frac{1}{|\mathcal{B}_B(t)|} \sum_{j \in \mathcal{B}_B(t)} \big(k_{\ell,h}(j) - \hat{\mu}_{\ell,h}(t)\big)\big(k_{\ell,h}(j) - \hat{\mu}_{\ell,h}(t)\big)^\top.$$

Using the Gaussian MGF, the expected global contribution is:

$$\widehat{A}_{\text{global}}(t) \;\approx\; |\mathcal{F}_W(t)| \cdot \exp\Big(\tau\, q_{\ell,h}(t)^\top \hat{\mu}_{\ell,h}(t) \;+\; \tfrac{1}{2}\tau^2\, q_{\ell,h}(t)^\top \hat{\Sigma}_{\ell,h}(t)\, q_{\ell,h}(t)\Big). \tag{17}$$

We then form the query-adaptive *local-mass fraction*:

$$\widehat{\pi}_{\ell,h}(t) \;=\; \frac{A_{\text{local}}(t)}{A_{\text{local}}(t) + \widehat{A}_{\text{global}}(t)}, \tag{18}$$

and aggregate over rows/examples on $\mathcal{D}$ to obtain a single head score (larger means more local):

$$s_{\ell,h} \;=\; \mathbb{E}_{(x,t)\in\mathcal{D}}\big[\widehat{\pi}_{\ell,h}(t)\big] \;\in [0, 1]. \tag{19}$$

B.3.5 Razor Attention (RAZOR)

Razor Attention (Tang et al., 2025) is a training-free KV cache compression method that aims to detects retrieval (long-context) heads, thus only keeping them in KV cache during decoding phase. Retrieval heads are detected via a synthetic repeated-token probe that isolates two patterns: *echo* (attending to the previous identical token) and *induction* (attending to the token that follows the current token in an earlier occurrence). The authors constructed an input of length $4K$ by repeating a random length-K token block four times. For target positions $m \in \{K, \ldots, 4K-1\}$, they define the *echo* and *induction* source indices:

$$j_{\text{echo}} = m - K, \qquad j_{\text{ind}} = m - 1 - K \;\; (\text{valid for } m \geq K+1), \tag{20}$$

and compute, for each layer ℓ and head h, the mean attention to these sources across the probe:

$$E_{\ell,h} \;=\; \mathbb{E}_m\big[\alpha_{\ell,h}(m, j_{\text{echo}})\big], \qquad I_{\ell,h} \;=\; \mathbb{E}_m\big[\alpha_{\ell,h}(m, j_{\text{ind}})\big]. \tag{21}$$

We convert these to across-head z-scores and into locality scores:

$$s_{\ell,h} \;=\; \max\big(z(E_{\ell,h}),\, z(I_{\ell,h})\big), \tag{22}$$

where smaller values indicate more local heads. We set K to one quarter of the sequence length of examples in D, to align with other methods.

B.3.6 Fisher-Weighted Locality (FISHER)

Kwon et al. (2022) propose a post-training structured pruning method that uses an empirical Fisher approximation of the loss to rank and select network components like attention heads and FFN blocks. We adpat this method by only considering attention heads and converting the Fisher-weighted attention mass into a per-head locality scores. Let $\alpha_{\ell,h}(t, j)$ be the attention probability from source j to target t (causal, $j \leq t$). We compute token-level cross-entropy $\mathcal{L}$ and its gradient in Fisher-style saliency on attention probabilities:

$$\Phi_{\ell,h}(t, j) \;=\; \Big(\tfrac{\partial \mathcal{L}}{\partial \alpha_{\ell,h}(t,j)} \cdot \alpha_{\ell,h}(t, j)\Big)^2, \tag{23}$$

then we bin by lag $d{=}t{-}j$ and normalize to a histogram:

$$F_{\ell,h}(d) \;=\; \sum_{(x,t)\in\mathcal{D}} \Phi_{\ell,h}\big(t,\, t{-}d\big), \qquad q_{\ell,h}(d) \;=\; \frac{F_{\ell,h}(d)}{\sum_{u\geq 0} F_{\ell,h}(u)}. \tag{24}$$

Given an SWA window W, the FISHER *locality* score is the Fisher-weighted mass inside the window:

$$s_{\ell,h} \;=\; \sum_{d=0}^{W-1} q_{\ell,h}(d) \;\in [0, 1], \tag{25}$$

where larger values indicate more *local* heads.

B.4 Calibration dataset

Our calibration set $\mathcal{D}$ contains 64 examples, each a 32k-token sequence, generated using in-house code. We verified that it has no content overlap with the NIAH test set used in the RULER benchmark (Hsieh et al., 2024) to ensure a fair evaluation. With our efficient BOSCH search mode, a single $\mathcal{S}(\mathcal{M}, z, \mathcal{D})$ call takes 2.2 seconds on Qwen3-1.7B-Base and 4.8 seconds on Qwen3-30B-A3B-Base, on a single node equipped with 8 A800 GPUs. Consequently, running both stages of BOSCH for a single configuration can take up to 4 hours on Qwen3-1.7B-Base and up to 14 hours on Qwen3-30B-A3B-Base, respectively. The same calibration dataset $\mathcal{D}$ is used to run baseline methods of § B.3. In comparison, the static baseline methods require less then one hour regardless of the model size using the same compute resources as our BOSCH. While BOSCH introduces higher overhead than baseline methods, the search is performed offline once per (model, SWA ratio), so the cost is amortized across all subsequent deployments and inference runs.

B.5 Continual Pretraining

To recover the full performance of the original self-attention model, we continually pre-train hybrid SWA models produced by our BOSCH method or the baselines on a small set of subsampled 2.5B tokens of Prolong (Gao et al., 2024) data. We perform tokenization and pack examples into fixed-length 32k token sequences. Models were trained on 2 GPU nodes, each equipped with 8 NVIDIA A800 cards with 80GB of memory. To accelerate pretraining, we use Fully Sharded Data Parallel (FSDP) (Zhao et al., 2023), mixed-precision training (Micikevicius et al., 2018), FlashAttention-2 (Shah et al., 2024), and Liger-Kernel (Hsu et al., 2024). We train all models on fully packed sequences of 32,768 tokens and set the maximum per-GPU batch size per model, ranging from 8 for the smallest Qwen3-1.7B-Base to 2 for the largest Qwen3-30B-A3B-Base. We further speed up training by adjusting the gradient-accumulation steps to achieve a total batch size of 2M tokens, as recommended by (Tang et al., 2024). For all models, we use the AdamW optimizer (Loshchilov and Hutter, 2017) with a cosine learning-rate scheduler, an initial learning rate of 1e-5, and a 10% warmup. Continual pretraining took approximately one day for Qwen3-1.7B-Base and 4 days for Qwen3-30B-A3B-Base.

B.6 Evaluation Benchmarks

We evaluate models on 2 benchmarks focusing on long-context reasoning and associative recall. We did not consider language understanding and modeling tasks such as ARC (Clark et al., 2018), PIQA (Bisk et al., 2020), and MMLU (Hendrycks et al., 2021), as they are regarded as short-context tasks where the total sequence length is smaller than our SWA window size.

For RULER (Hsieh et al., 2024), we include three NIAH-single (`niah_single_{1,2,3}`) and three NIAH-multikey (`niah_multikey_{1,2,3}`) subtasks evaluated at 4k, 8k, 16k, and 32k context lengths. For each length, we report the unweighted average across the three NIAH-single subtasks and across the three NIAH-multikey subtasks. The final **NIAH** average score is the mean over the 6 subtasks scores.

For **LongBench** (Bai et al., 2024) (v1), we follow the authors' categorizations: Single-document QA (MultiFieldQA-en, MultiFieldQA-zh, NarrativeQA, Qasper), Multi-document QA (2WikiMQA, HotpotQA, MuSiQue, DuReader), Summarization (GovReport, MultiNews, QMSum, VCSum, SAMSum), Few-shot learning (TREC, TriviaQA, LSHT), Synthetic Tasks (PassageRetrieval-en, PassageRetrieval-zh, PassageCount), and Code Completion (LCC, RepoBench-P). We add an extra category, which we call *Math Reasoning*, consisting of a single task: GSM8K (Cobbe et al., 2021). The **LongBench** score is the unweighted average over seven categories (the 6 LongBench categories plus Math Reasoning).

C Analysis

C.1 BOSCH Head Analysis

We conduct additional experiments to validate whether the findings of § 4.3 on BOSCH heads selection is correlated with model performance. Let $\mathcal{A}_\rho$ denote the set of selected heads at ratio ρ. For two small and larger ratios $\rho_s < \rho_l$, define $\mathcal{A}_{\rho_s}$ and $\mathcal{A}_{\rho_l}$ to be the heads selected by these two ratios, respectively. Let

$$\Delta\mathcal{A}_{s,l} = \mathcal{A}_{\rho_s} \setminus \mathcal{A}_{\rho_l}$$

be the set of heads that exist at ratio ρ_s but not at ρ_l; using this set, we randomly select the same number of heads from $\mathcal{A}_{\rho_l}$ and replace them with the heads in $\Delta\mathcal{A}_{s,l}$, leading to $\mathcal{A}'_{\rho_l}$.

Model	NIAH						LongBench					
	$0.25 \rightarrow 0.5$		$0.5 \rightarrow 0.75$		$0.75 \rightarrow 0.875$		$0.25 \rightarrow 0.5$		$0.5 \rightarrow 0.75$		$0.75 \rightarrow 0.875$	
	$\mathcal{A}_{\rho_l}$	$\mathcal{A}'_{\rho_l}$	$\mathcal{A}_{\rho_l}$	$\mathcal{A}'_{\rho_l}$	$\mathcal{A}_{\rho_l}$	$\mathcal{A}'_{\rho_l}$	$\mathcal{A}_{\rho_l}$	$\mathcal{A}'_{\rho_l}$	$\mathcal{A}_{\rho_l}$	$\mathcal{A}'_{\rho_l}$	$\mathcal{A}_{\rho_l}$	$\mathcal{A}'_{\rho_l}$
1.7B	78.3	57.1	58.0	47.1	30.3	19.7	32.1	27.4	26.0	23.0	23.6	17.0
8B	90.3	79.5	72.7	68.7	42.5	20.7	41.8	34.3	31.6	29.2	28.6	24.3
14B	94.0	87.9	83.6	73.9	47.2	41.1	47.0	40.4	38.0	34.0	36.1	29.9
30B	86.3	67.2	50.2	36.8	26.9	15.7	36.0	28.4	24.8	22.3	19.3	14.7

Table 5: Zero-shot average scores on the NIAH and LongBench benchmarks for BOSCH across 4 Qwen3 models under 3 SWA ratios $\rho_l \in \{0.5, 0.75, 0.875\}$. The left side of $\rightarrow$ corresponds to ρ_s, while the right side corresponds to ρ_l. $\mathcal{A}'_{\rho_l}$ indicates results obtained after replacing a randomly selected subset of heads at a given configuration (e.g., $\rho_l = 0.75$) with heads that did not appear in a smaller configuration (e.g., $\rho_s = 0.5$). Results in the $\mathcal{A}'_{\rho_l}$ columns are averaged over 3 runs with different seeds. The $\mathcal{A}_{\rho_l}$ column shows the results for the originally selected SWA heads set.

Method	single				multikey				Avg.			
	1.7B	**8B**	**14B**	**30B**	**1.7B**	**8B**	**14B**	**30B**	**1.7B**	**8B**	**14B**	**30B**
64k												
Original	47.9	49.3	49.1	49.3	23.5	44.7	45.9	45.7	35.7	47.0	47.5	47.5
INTR	2.6	2.9	2.6	17.3	1.7	1.8	1.3	2.5	2.2	2.4	2.0	9.9
B-layer	23.7	18.1	12.1	19.5	8.2	1.6	2.7	3.8	16.0	9.9	7.4	11.7
FISHER	27.7	37.0	37.0	23.9	7.5	8.7	13.3	5.7	17.6	22.9	25.2	14.8
BOSCH	**32.0**	**41.3**	**43.7**	**38.5**	**15.0**	**13.9**	**23.6**	**11.1**	**23.5**	**27.6**	**33.7**	**24.8**
128k												
Original	21.6	24.3	24.4	22.4	6.1	14.7	18.1	16.5	13.8	19.5	21.3	19.4
INTR	1.5	1.7	1.5	6.1	1.0	1.1	0.9	1.5	1.3	1.4	1.2	3.8
B-layer	4.5	9.9	2.2	8.7	2.3	1.2	1.1	1.5	3.4	5.6	1.7	5.1
FISHER	6.9	9.1	15.3	8.8	1.8	1.1	6.4	3.1	4.4	5.1	10.9	6.0
BOSCH	**11.7**	**10.0**	**18.7**	**15.3**	**5.1**	**2.4**	**6.5**	**4.9**	**8.4**	**6.2**	**12.6**	**10.1**

Table 6: Detailed zero-shot performance on the NIAH benchmark at *64k* and *128k* for the original models and four SWA hybridization methods (with SWA ratio $\rho = 0.75$) applied to four Qwen3 models (1.7B, 8B, 14B, 30B). Zero-shot length extrapolation is performed using a YaRN factor of 2 and 4 for 64k and 128k, respectively. The **Avg.** block reports, for each model, the mean of its single and multikey scores. The highest score under each configuration is highlighted in bold.

Table 5 shows the zero-shot average NIAH and LongBench performance for BOSCH when using three randomly sampled $\mathcal{A}'_{\rho_l}$ head sets based on $\mathcal{A}_{\rho_l}$ (mean of 3 runs with different $\mathcal{A}'_{\rho_l}$). Results for $\mathcal{A}_{\rho_l}$ are reported for reference, while the detailed NIAH and LongBench scores are presented in Table 19 and Table 20, respectively. We notice a significant drop in performance across all models, benchmarks, and ratios when using $\mathcal{A}'$ instead of $\mathcal{A}$. For instance, on NIAH, the gap varies between 6% (14B, $\rho_s \in \{0.5, 0.875\}$) and 20% (14B, $\rho_s = 0.5$; 8B, $\rho_s = 0.875$). These observations further indicate that an SWA hybridization method should not only rely on pre-hybridization head locality identification, but also explicitly account for the impact of head entanglement after hybridization.

C.2 Length Extrapolation Ablation

We evaluate on the NIAH benchmark using sequences that exceed each model's native maximum context length to study how robust SWA hybridization methods are to longer sequence lengths. For all models, we perform zero-shot extrapolation of the context length by replacing the default RoPE positional encoding (Su et al., 2024) with the Yarn context window extension technique (Peng et al.,

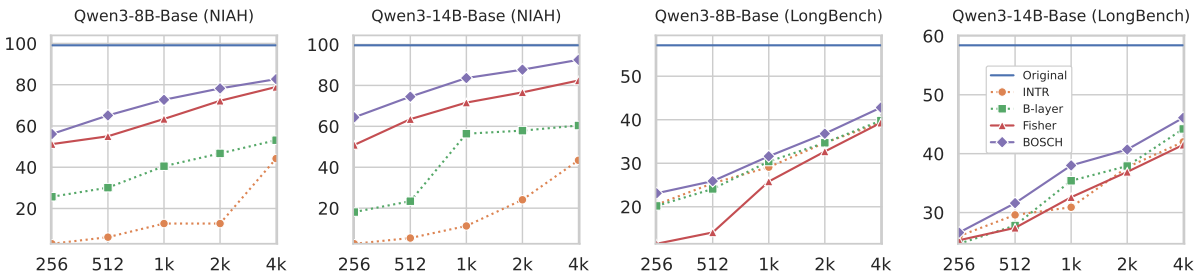

Figure 4: NIAH (first 2 plots) and LongBench (last 2 plots) zero-shot average performances (y-axis) for Qwen3-8B-Base and Qwen3-14B-Base models as a function of varying the SWA window size (x-axis). We report scores for two layer-level and two head-level SWA hybridization methods using $\rho = 0.75$. The performance of the full-attention original models (without additional training) is shown as a straight line with no markers.

2024). More precisely, we set the YaRN factor to 2.0 and 4.0 for inference sequence lengths of *64k* and *128k*, respectively, without any additional finetuning.

Table 6 reports NIAH zero-shot performance at *64k* and *128k* sequence lengths for the original model, as well as for the four SWA hybridization methods using with SWA ratio $\rho = 0.75$. We find that BOSCH remains the strongest SWA hybrid across all model sizes and both sequence lengths, consistently outperforming INTR, *B-layer*, and FISHER. We also observe that the relative ranking among methods is largely stable compared to the 32k setting, while BOSCH's margin over layer-level heuristics becomes more significant as context length increases. This suggests that BOSCH is not only effective within the native 32k window, but also transfers reasonably well under zero-shot length extrapolation to longer sequences such as 64k and 128k.

C.3 SWA Windows Size Ablation

We conduct an ablation study by varying the SWA window size to better understand how the methods in our study generalize when evaluated with a window size that was not used during search or ranking. Figure 4 shows the NIAH and LongBench zero-shot average scores of 4 methods for Qwen3-8B-Base and Qwen3-14B-Base when varying the SWA window size with values $\{256, 512, 1024, 2048, 4096\}$ (x-axis), while search and ranking are performed with a fixed window size of 1024 and SWA ratio $\rho = 0.75$. Results for the Qwen3-1.7B-Base and Qwen3-30B-A3B-Base models, as well as detailed results for all models, are reported in Table 17 and Table 18 for NIAH and LongBench, respectively.

Across all tested window sizes and models, BOSCH systematically continues to outperform the remaining methods on both benchmarks. Moreover, the relative ranking of the baselines largely matches the one observed at the search window size ($W = 1024$), with only few expections. In addition, we observe that increasing the window size generally leads to an approximately monotonic performance improvements under most setting. This suggests that the quality of a head-selection scheme is fairly stable and predictable when the SWA windows size change at inference time evaluation, and that BOSCH's advantage does not limited to the window size used during search.

C.4 Latency and Memory Analysis

Figure 5 shows latency and memory usage when comparing the original Qwen3-8B-Base model with BOSCH SWA hybrid models with $\rho \in \{0.25, 0.5, 0.75, 0.875\}$ and a window size of 1024. Specifically, we measure prefill and decoding throughput (tokens/s) and the p90 peak memory usage (in GB) when running inference on sequences of varying length (64k, 256k, 512k, and 1M tokens). We run systematic experiments for all models with a batch size of 1 and a streaming context size of 8192 to avoid early out-of-memory (OOM) errors on ultra-long sequences. Missing data points (e.g., for lengths >256k in the original model) indicate that the model ran out of memory (OOM).

Although $\rho = 0.25$ leads to no significant performance drop, it also does not provide efficiency gains: at this ratio, the model still goes OOM beyond 256k tokens, just like the original model despite delivering better latency. In contrast, $\rho = 0.5$ yields improved latency with a moderate performance drop, but it is still not suitable for ultra-long contexts beyond 512k tokens, unlike $\rho = 0.75$ and $\rho = 0.875$, which can handle longer sequences.

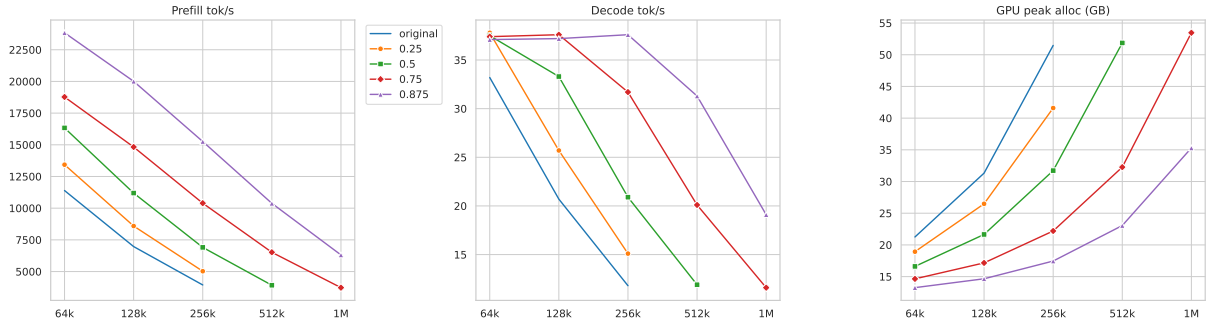

Figure 5: Latency and memory statistics comparing the original Qwen3-8B-Base model with SWA hybrid variants at different SWA ρ ratios using BOSCH. The left plot shows prefill throughput (tokens/s), the middle plot shows decoding throughput in the same units, and the right plot shows p90 memory allocation (GB). These statistics are measured with input prompts of lengths ranging from 64 to 1M tokens (x-axis). Missing data points correspond to runs where the model encountered out-of-memory (OOM) error.

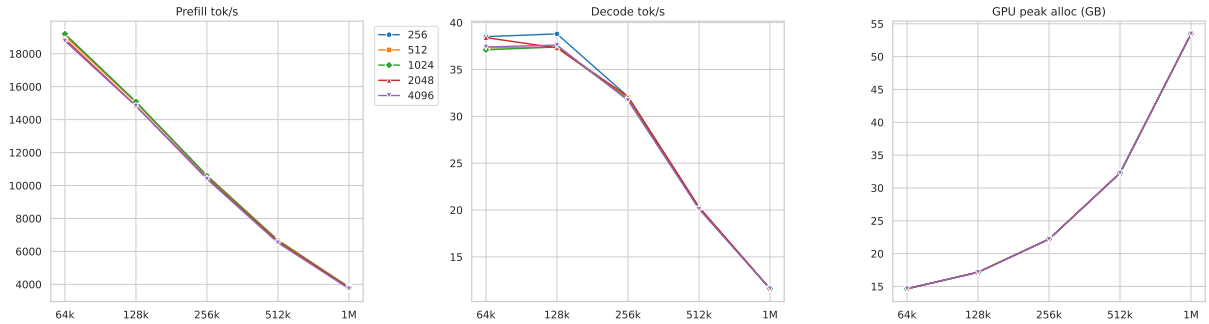

Figure 6: Latency and memory statistics of BOSCH $\rho = 0.75$ Qwen3-8B-Base SWA hybrid models when varying the SWA window size between 256 and 4096. Notation is the same as in Figure 5.

Considering the performance gap between the latter two, $\rho = 0.75$ appears to offer a better trade-off between performance and efficiency. Figure 6 presents the same statistics as Figure 5, but for the BOSCH Qwen3-8B-Base SWA hybrid model with $\rho = 0.75$ while varying the SWA window size across $\{256, 512, 1024, 2048, 4096\}$. Despite minor variations for short sequences, we observe that all metrics are almost identical across window sizes when scaling to long sequences. This is mainly because the computational complexity of SWA is effectively constant with respect to the window size, making sequence length the primary driver of complexity. This is encouraging, as it indicates that one can use larger window sizes to benefit from performance gains (see § C.3) without sacrificing latency.

D Results

Method	single																multikey																Avg.			
	4k				8k				16k				32k				4k				8k				16k				32k							
	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875
Original	100.0				99.9				100.0				100.0				98.9				94.9				84.1				63.9				92.7			
Layer-level Heuristics																																				
RAND	95.2	71.9	32.4	32.0	87.5	58.8	15.7	15.6	76.5	22.3	6.5	7.5	65.4	21.4	2.9	2.8	71.3	51.7	19.9	17.8	47.1	23.8	11.1	10.9	32.9	14.5	4.6	4.5	23.0	7.6	1.5	1.5	66.9	38.0	12.7	12.6
BME	85.9	32.2	32.2	32.2	69.7	15.6	15.5	15.5	60.0	7.5	7.5	7.5	43.9	2.7	2.8	2.8	33.7	15.9	19.5	19.9	16.1	9.6	10.4	10.5	11.1	4.1	3.7	4.3	5.3	1.3	1.1	1.3	40.0	11.5	11.1	11.2
INTR	85.9	43.7	32.1	32.2	76.6	23.0	15.6	15.5	78.5	17.6	7.5	7.4	74.2	10.1	2.8	2.8	87.9	35.3	20.9	21.4	65.7	12.5	11.4	11.8	52.3	5.2	4.5	4.4	35.0	2.2	1.7	1.5	74.9	19.1	12.8	12.5
Head-level Static Methods																																				
DCAM	32.6	16.2	15.5	15.1	99.4	88.5	46.6	18.1	98.2	9.7	7.8	7.9	96.2	5.5	2.9	3.0	91.1	19.6	14.4	15.5	73.1	7.7	7.7	7.1	48.9	4.5	3.3	2.7	38.8	2.4	0.7	0.9	80.7	12.3	10.6	10.6
APL	32.3	15.6	15.6	15.0	23.3	15.6	15.6	15.0	18.0	7.5	7.5	7.9	11.5	2.8	2.8	3.0	44.9	31.2	15.6	23.0	25.5	12.3	7.4	10.9	11.9	4.8	3.1	4.1	4.5	1.8	0.8	1.3	22.9	13.5	10.6	12.2
PROXY	99.2	59.3	32.2	32.5	97.1	43.8	16.8	15.1	65.9	21.1	7.6	7.9	50.3	12.3	2.9	3.0	50.9	29.3	16.7	20.1	23.2	13.9	9.2	9.6	13.8	7.3	3.5	4.1	6.5	3.7	1.1	1.2	50.9	23.8	11.3	11.7
QADA	99.9	92.9	66.3	33.5	**100.0**	88.5	46.6	18.1	99.3	66.9	52.3	15.2	90.9	50.3	35.7	19.1	93.7	71.7	41.3	20.5	76.5	48.3	28.5	10.1	63.4	39.5	24.3	6.8	44.6	23.7	16.3	8.5	83.5	60.2	38.9	16.5
RAZOR	**100.0**	99.1	82.5	35.6	**100.0**	99.5	82.1	24.1	**99.9**	99.3	65.8	31.4	99.6	96.3	69.4	**24.5**	95.3	43.3	28.8	22.8	80.7	33.0	21.3	13.7	62.7	25.9	17.2	13.5	45.5	22.2	15.6	**9.3**	85.5	64.8	47.8	21.9
FISHER	**100.0**	**99.7**	**87.3**	32.3	99.9	**99.6**	79.5	14.8	97.9	**99.4**	72.9	8.0	**99.7**	**98.8**	52.9	3.0	95.1	80.7	40.1	16.5	85.6	55.0	28.2	7.9	68.2	49.7	21.1	2.9	51.5	28.5	12.8	1.0	87.2	76.4	49.4	10.8
Ours																																				
BOSCH	99.8	88.2	76.4	**62.4**	99.9	85.1	77.9	**49.1**	99.6	86.3	72.7	**37.8**	99.5	12.5	63.1	12.5	**98.3**	76.9	**63.5**	**37.7**	**89.9**	**74.1**	**45.4**	**20.2**	**81.7**	**66.9**	**33.0**	**17.0**	**66.7**	**57.1**	**31.3**	5.5	**91.8**	**78.3**	**58.0**	**30.3**
-single	98.4	67.5	55.5	37.6	96.5	78.6	44.1	37.1	88.3	78.4	25.0	24.9	96.7	73.4	20.2	8.9	94.6	74.4	37.1	16.1	85.9	52.7	25.5	16.8	77.7	42.3	9.9	13.6	56.9	35.1	5.7	4.3	86.9	62.8	27.9	19.9
-multi	**100.0**	83.6	79.5	42.8	**100.0**	89.6	80.1	31.1	99.9	81.9	73.7	13.8	99.7	74.4	**70.1**	17.6	96.9	**85.5**	52.2	20.7	88.1	60.3	40.9	12.3	74.7	55.1	30.9	11.0	55.3	48.1	30.7	7.9	89.5	71.2	57.3	19.6
-layer	91.7	91.9	83.1	15.1	99.0	91.9	**83.1**	15.1	98.8	90.9	**78.6**	7.9	97.5	80.3	55.1	3.0	96.5	77.3	37.7	22.3	91.1	49.3	20.1	11.4	79.7	40.7	20.7	4.9	58.3	32.9	13.7	1.5	90.0	69.4	49.4	12.3

Table 7: Detailed zero-shot scores on the NIAH benchmark for SWA hybridization methods for Qwen3-1.7B-Base under 4 SWA ρ ratios. The highest score under each configuration is highlighted in bold.

Method	Single-document QA				Multi-document QA				Summarization				Few-shot learning				Synthetic Tasks				Code Completion				Math Reasoning				Avg.			
	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875
Original	39.8				30.9				26.0				67.4				32.7				48.2				72.6				45.4			
Layer-level Heuristics																																
RAND	27.0	17.6	13.1	14.9	15.7	10.0	9.4	11.1	23.3	19.1	16.3	16.4	56.1	46.3	36.6	36.4	4.2	3.6	2.6	3.2	49.8	48.5	**47.2**	44.9	60.0	**54.4**	38.1	34.5	33.7	28.5	23.3	23.1
BME	23.1	12.7	12.9	14.1	13.8	11.3	11.1	10.2	21.1	14.8	15.7	15.9	58.1	36.8	37.8	37.0	3.2	2.9	2.8	2.5	50.0	32.3	34.6	36.7	51.0	36.8	29.9	30.3	31.5	21.1	20.7	20.9
INTR	29.8	19.3	13.6	13.2	13.1	10.8	11.8	12.4	24.4	19.4	15.6	15.5	59.7	44.7	39.0	36.6	7.8	4.1	3.6	3.0	53.3	46.2	26.8	22.5	63.9	53.3	28.4	**35.0**	36.0	28.2	19.8	19.7
Head-level Static Methods																																
DCAM	29.5	12.3	8.0	9.6	20.4	6.7	6.9	6.8	23.2	18.0	17.0	14.0	61.8	45.2	29.8	34.0	3.6	4.1	2.1	2.5	28.2	25.7	33.0	28.5	60.7	45.4	32.2	29.3	32.5	22.5	18.4	17.8
APL	30.6	10.3	6.9	7.8	21.5	5.5	5.5	7.0	25.2	19.4	14.4	13.4	62.0	39.9	22.2	25.6	7.1	2.6	1.5	2.6	43.9	**49.8**	32.0	33.6	**67.2**	53.6	31.8	26.4	36.8	25.9	16.3	16.6
PROXY	30.0	12.9	10.6	10.7	19.3	8.9	8.9	9.8	23.1	18.7	14.9	13.4	58.8	41.7	36.3	26.8	5.7	2.0	2.3	3.0	44.1	45.5	35.0	37.4	59.8	40.6	29.9	26.8	34.4	24.3	19.7	18.3
QADA	27.2	20.3	16.9	15.9	21.4	15.0	6.5	**12.9**	23.5	21.6	19.4	15.8	59.7	53.5	**48.5**	**41.6**	6.9	**5.2**	**4.4**	2.9	**54.8**	34.8	**44.3**	31.6	61.5	46.1	34.9	31.5	36.4	28.1	25.0	21.7
RAZOR	31.2	24.2	14.4	14.0	24.8	17.5	10.3	11.0	24.1	20.7	13.9	15.4	63.1	55.9	48.1	41.6	10.7	4.0	3.7	3.1	34.7	31.3	16.6	25.9	57.4	41.5	31.5	27.7	35.1	27.9	19.8	19.8
FISHER	34.7	28.6	16.2	**15.7**	28.1	20.5	11.8	10.8	24.3	21.6	17.5	14.8	**66.5**	59.4	46.1	40.8	5.0	4.2	3.0	2.9	34.3	28.1	27.9	21.0	60.1	45.3	27.9	30.3	36.1	29.7	21.5	19.5
Ours																																
BOSCH	35.1	**29.2**	**18.5**	12.5	**29.2**	**22.7**	**13.5**	10.2	23.7	**22.6**	19.5	**17.0**	62.0	**60.1**	48.2	38.3	4.9	4.4	3.6	3.5	43.7	32.9	38.9	**50.0**	67.1	53.1	**40.2**	34.0	**38.0**	**32.1**	**26.0**	**23.6**
-single	30.7	11.9	9.3	12.5	17.4	5.6	7.9	12.9	24.4	16.3	14.0	12.9	55.8	48.7	36.9	38.5	**18.2**	4.2	3.5	3.4	44.5	49.6	27.4	25.6	63.1	47.8	37.0	29.0	36.3	26.3	19.4	18.5
-multi	32.2	18.4	8.8	11.2	23.9	12.7	5.3	3.6	23.9	19.4	16.1	16.5	62.1	50.3	43.5	39.3	9.6	4.8	2.4	3.6	39.6	48.0	39.3	32.4	61.0	52.8	36.9	30.5	36.0	29.5	21.8	20.3
-layer	**35.2**	18.4	18.2	12.7	24.0	14.4	12.3	10.4	24.8	17.7	**19.9**	14.2	64.1	51.6	43.1	31.9	11.2	2.2	3.9	2.4	35.9	34.7	36.1	33.7	64.4	52.6	39.5	28.7	37.1	27.4	24.7	19.1

Table 8: Detailed zero-shot scores on the LongBench benchmark for SWA hybridization methods for Qwen3-1.7B-Base under 4 SWA ρ ratios. The highest score under each configuration is highlighted in bold.

Method	single																multikey																Avg.			
	4k				8k				16k				32k				4k				8k				16k				32k							
	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875
Original	100.0				100.0				100.0				99.9				99.5				99.6				99.1				94.5				99.1			
Layer-level Heuristics																																				
RAND	68.5	37.0	32.0	32.2	60.4	17.6	15.8	15.6	53.6	8.1	7.5	7.5	49.9	3.3	2.8	2.8	58.8	34.6	26.5	27.6	38.5	14.4	11.7	12.8	22.7	6.0	4.5	5.3	15.0	2.1	1.6	1.7	45.9	15.4	12.8	13.2
BME	68.9	32.3	32.1	32.3	51.3	15.8	15.6	15.6	39.7	7.9	7.4	7.5	25.5	3.7	2.8	2.8	36.6	22.3	22.8	24.7	14.5	11.0	11.5	12.2	7.3	4.6	4.2	4.7	2.5	1.5	1.3	1.7	30.8	12.4	12.2	12.7
INTR	92.9	40.2	32.3	32.1	92.3	21.9	15.6	15.6	88.1	17.1	7.6	7.5	79.5	9.3	2.9	2.7	78.6	37.1	23.9	24.1	64.2	15.3	12.5	12.5	57.6	7.5	4.7	4.8	30.1	3.5	1.7	1.8	72.9	19.0	12.7	12.6
Head-level Static Methods																																				
DCAM	**100.0**	99.9	49.8	34.3	**100.0**	99.6	36.0	16.9	**100.0**	97.9	19.3	8.5	99.5	92.1	16.6	3.2	98.0	96.1	35.6	24.8	96.3	84.3	15.4	11.3	94.8	64.9	5.3	5.1	77.9	40.5	2.5	1.2	95.8	84.4	22.6	13.2
APL	99.8	95.3	32.3	31.9	99.8	90.5	15.6	15.6	99.7	82.9	7.5	7.5	99.7	66.3	2.8	2.8	**99.5**	87.3	26.5	22.4	**99.5**	59.5	12.7	11.0	98.3	38.7	4.9	4.3	88.2	20.1	1.6	1.5	98.0	67.6	13.0	12.1
PROXY	**100.0**	99.1	61.0	33.0	99.9	99.3	53.3	16.5	99.9	96.5	42.3	7.9	99.6	67.8	14.1	2.8	**99.5**	92.5	40.0	29.6	98.7	61.0	21.5	13.8	96.4	36.3	14.5	5.9	88.3	20.0	6.5	1.9	97.8	71.6	31.6	13.9
QADA	**100.0**	99.7	77.1	33.0	99.9	99.5	81.5	17.3	99.7	95.8	66.0	12.1	**99.9**	91.7	49.0	4.3	98.2	88.5	42.9	20.5	96.7	63.7	25.6	10.5	92.3	44.0	16.9	4.7	74.8	24.2	10.3	1.5	95.2	75.9	46.1	13.0
RAZOR	**100.0**	**100.0**	96.7	68.3	99.9	99.9	**95.9**	56.2	99.9	99.9	95.4	53.0	**99.9**	**98.4**	71.9	10.7	98.6	95.4	70.9	41.1	96.8	77.7	41.8	**21.7**	91.1	55.7	31.7	**17.1**	66.9	32.1	14.8	3.1	94.1	82.4	64.9	33.9
FISHER	**100.0**	99.9	93.2	68.1	**100.0**	**100.0**	86.8	32.2	**100.0**	**100.0**	77.4	47.2	**99.9**	97.9	71.5	17.3	98.5	96.3	72.2	30.3	96.2	87.0	50.4	15.3	89.3	78.8	31.9	14.9	69.6	54.1	24.2	**6.7**	94.2	89.3	63.4	29.0
Ours																																				
BOSCH	**100.0**	99.0	**98.7**	**81.3**	**100.0**	99.0	95.3	**70.1**	99.6	98.1	**97.1**	**68.6**	99.2	96.3	**84.9**	**34.1**	99.4	95.9	**86.8**	**44.3**	99.3	89.4	**51.2**	21.3	**98.5**	**82.9**	**40.2**	15.0	**94.9**	**61.9**	**27.4**	5.3	**98.9**	**90.3**	**72.7**	**42.5**
-single	99.5	99.0	75.8	31.9	98.3	98.0	60.3	15.6	94.6	97.5	49.2	7.3	92.4	96.9	32.7	2.8	96.5	95.9	57.8	23.0	93.3	85.3	29.6	11.6	89.3	79.0	17.3	4.8	64.3	59.4	12.8	1.7	91.0	88.9	41.9	12.3
-multi	99.8	96.1	94.5	44.2	99.7	91.5	84.5	21.3	99.3	77.9	67.1	18.6	97.6	61.2	40.3	3.7	**99.5**	74.7	51.7	34.3	98.7	38.5	23.5	13.1	97.6	24.2	18.0	9.6	88.3	15.8	8.5	2.1	97.6	60.0	48.5	18.4
-layer	99.2	99.9	79.3	32.1	99.1	99.5	74.5	15.6	98.5	98.9	65.2	7.4	97.2	92.0	46.1	2.7	97.3	**98.7**	28.5	25.7	94.1	**94.3**	15.5	12.8	87.1	81.0	9.5	5.5	73.4	53.4	5.7	1.9	93.2	89.7	40.6	13.0

Table 9: Detailed zero-shot scores on the NIAH benchmark for SWA hybridization methods for Qwen3-8B-Base under 4 SWA ρ ratios. The highest score under each configuration is highlighted in bold.

Method	Single-document QA				Multi-document QA				Summarization				Few-shot learning				Synthetic Tasks				Code Completion				Math Reasoning				Avg.			
	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875
Original	45.5				41.9				26.8				70.3				63.0				66.6				85.8				57.1			
Layer-level Heuristics																																
RAND	25.8	14.6	7.5	9.8	20.4	11.6	5.3	7.4	24.5	18.7	16.6	16.5	62.9	50.0	38.1	36.5	12.8	3.8	3.5	3.0	52.5	43.2	32.9	33.1	82.3	78.8	65.2	64.0	40.2	31.5	24.1	24.3
BME	19.3	10.8	14.4	10.3	15.4	9.8	14.8	12.0	21.0	19.2	20.6	18.8	50.7	42.6	41.4	38.2	4.0	3.8	2.9	2.7	29.9	26.3	45.0	42.0	76.3	58.2	62.9	61.3	31.0	24.4	28.9	26.5
INTR	29.1	18.9	18.1	11.9	26.1	12.5	**15.1**	8.2	25.4	22.5	20.0	15.6	63.9	48.7	40.7	36.3	12.1	4.1	2.4	**3.9**	**64.2**	53.7	44.1	26.4	82.4	78.8	63.4	62.2	43.3	34.2	29.1	23.5
Head-level Static Methods																																
DCAM	38.5	31.5	17.2	13.7	34.2	20.2	14.2	11.6	25.3	20.8	16.9	17.7	68.4	56.8	38.6	34.3	19.8	5.9	3.9	3.6	33.9	23.6	28.4	27.6	81.3	71.9	62.9	59.4	43.0	33.0	26.0	24.0
APL	40.9	22.8	15.1	10.8	**36.0**	13.2	13.2	11.2	25.7	20.8	18.4	16.4	70.1	59.3	39.2	36.3	43.3	7.5	3.4	3.5	58.6	38.0	27.4	29.1	83.2	81.5	60.2	58.3	51.1	34.7	25.3	23.7
PROXY	37.8	23.2	16.0	14.0	35.4	11.8	7.0	9.6	25.3	22.2	21.0	**20.9**	66.6	52.7	48.2	41.4	46.1	8.3	4.1	3.3	60.3	37.6	38.0	37.4	82.2	79.9	70.9	64.7	50.5	33.7	29.3	27.3
QADA	34.2	20.2	8.9	9.2	22.3	11.5	6.7	7.5	23.2	22.7	12.3	13.9	69.4	60.7	37.8	38.6	12.4	5.0	3.5	3.2	38.2	17.8	21.0	26.8	81.3	72.1	57.9	63.9	40.1	30.0	21.2	23.3
RAZOR	32.8	13.8	15.3	10.6	31.5	7.1	10.6	11.4	20.7	14.4	16.9	18.1	67.5	57.7	47.5	**46.4**	18.2	7.2	3.4	2.8	35.7	28.3	23.0	30.8	82.1	75.7	63.0	**67.2**	41.2	29.2	25.7	26.7
FISHER	33.9	26.4	12.8	8.8	19.3	15.1	6.0	6.7	24.4	22.7	13.6	13.9	**70.3**	**64.4**	49.2	41.5	19.8	11.7	5.2	2.4	36.8	25.2	31.0	32.2	81.6	77.6	62.8	59.8	40.9	34.7	25.8	23.6
BOSCH																																
BOSCH	37.1	**32.3**	**20.2**	**16.5**	33.5	**23.5**	14.8	**13.5**	24.0	**25.6**	**21.6**	19.6	66.5	61.9	**51.4**	39.7	**61.3**	**12.6**	**6.4**	2.6	59.8	**54.0**	32.6	**43.2**	**83.5**	**82.4**	74.0	65.1	**52.2**	**41.8**	**31.6**	**28.6**
-single	29.6	27.2	8.4	11.8	31.5	17.5	9.7	10.6	23.0	22.9	15.4	16.6	67.3	53.9	44.4	44.5	31.0	7.4	4.3	3.2	52.7	48.2	30.2	29.8	83.2	77.9	70.9	63.0	45.5	36.4	26.2	27.4
-multi	40.2	25.2	13.8	11.6	34.1	17.9	8.8	12.1	25.7	23.2	21.4	18.3	67.5	54.9	36.2	42.3	23.8	2.7	4.4	2.9	53.8	35.0	28.8	38.9	81.7	79.5	72.9	62.5	46.7	34.1	26.6	26.9
-layer	**41.8**	24.4	7.9	8.3	32.4	19.0	7.5	9.5	**25.9**	23.5	19.0	15.9	66.1	58.2	46.5	38.2	23.2	8.3	3.2	2.8	61.7	41.0	**54.2**	28.4	82.6	79.5	**74.5**	51.1	47.7	36.3	30.4	22.0

Table 10: Detailed zero-shot scores on the LongBench benchmark for SWA hybridization methods for Qwen3-8B-Base under 4 SWA ρ ratios. The highest score under each configuration is highlighted in bold.

Method	single																multikey																Avg.			
	4k				8k				16k				32k				4k				8k				16k				32k							
	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875
Original	100.0				99.9				100.0				100.0				99.8				99.7				99.3				98.2				99.6			
Layer-level Heuristics																																				
RAND	99.9	86.4	42.1	32.3	97.3	59.1	22.8	15.6	92.9	47.8	11.7	7.5	82.9	31.7	8.8	2.8	96.2	54.0	30.7	26.9	74.6	22.4	13.5	12.5	53.0	9.4	5.5	4.9	35.6	3.0	1.6	1.6	79.0	39.2	17.1	13.0
BME	98.5	31.3	32.3	32.3	82.5	15.5	15.6	15.6	43.7	7.1	7.4	7.5	34.1	2.7	2.8	2.8	41.3	22.1	20.6	21.9	20.8	11.1	11.0	11.4	7.1	3.7	4.2	4.6	2.1	1.3	1.6	1.5	41.3	11.8	11.9	12.2
INTR	65.8	32.2	32.4	32.3	46.6	15.4	15.7	15.6	33.4	7.5	7.5	7.5	22.5	2.8	2.9	2.8	68.7	32.8	17.7	17.8	40.7	14.5	9.1	9.7	33.7	5.9	3.5	3.9	17.7	2.1	1.3	1.5	41.1	14.1	11.3	11.4
Head-level Static Methods																																				
DCAM	**100.0**	99.5	57.8	32.9	99.9	99.3	47.7	16.0	**100.0**	99.2	27.9	7.9	98.8	97.9	28.9	4.5	98.9	91.5	43.7	25.5	98.1	84.8	20.9	12.1	95.9	69.3	9.2	5.1	83.4	54.3	5.7	2.0	96.9	87.0	30.2	13.2
APL	99.9	**100.0**	32.2	32.1	99.9	99.9	15.6	15.3	**100.0**	**99.9**	7.5	7.5	**100.0**	98.7	2.8	2.8	99.7	98.6	33.1	26.3	98.9	90.7	14.0	12.5	98.9	77.1	5.8	4.8	95.1	65.6	2.0	1.7	99.0	91.3	14.1	12.9
PROXY	**100.0**	99.3	87.6	32.3	99.9	96.5	67.6	15.6	**100.0**	84.8	59.0	7.4	**100.0**	81.7	21.0	2.8	99.5	88.1	52.0	31.7	97.7	45.7	25.7	13.2	95.5	31.3	19.3	5.7	86.1	21.1	9.4	2.0	97.3	68.6	42.7	13.8
QADA	99.9	98.7	50.9	39.6	98.2	96.0	26.3	29.6	99.7	86.3	10.5	17.5	88.9	84.2	10.1	5.4	88.7	75.7	31.2	28.1	77.1	54.5	14.6	15.2	66.9	39.3	6.0	5.7	51.7	31.9	3.5	2.1	83.9	70.8	19.1	17.9
RAZOR	99.9	**100.0**	95.7	**80.1**	**100.0**	**100.0**	94.0	53.6	**100.0**	**99.9**	**97.6**	62.9	**100.0**	**99.8**	83.5	37.7	99.7	98.0	78.9	**60.9**	99.0	80.8	47.3	**34.8**	98.8	70.2	42.0	**29.2**	93.7	54.9	31.2	**16.2**	98.9	88.0	71.3	46.9
FISHER	99.9	**100.0**	**96.1**	70.4	99.9	**100.0**	**96.1**	67.0	**100.0**	**99.9**	93.9	44.3	99.9	99.7	**89.7**	37.5	99.4	98.0	77.3	43.1	98.7	94.5	50.1	25.9	98.2	86.2	38.8	17.1	94.0	71.5	30.9	14.1	98.8	93.7	71.6	39.9
BOSCH																																				
BOSCH	**100.0**	99.9	94.5	77.3	99.9	99.3	95.9	**73.9**	**100.0**	97.8	97.5	**66.9**	99.8	92.2	78.5	47.9	99.4	**99.1**	**95.0**	47.2	99.1	**97.1**	**78.9**	30.9	**99.0**	**92.7**	**71.3**	19.6	**96.7**	73.8	**56.9**	13.7	**99.2**	**94.0**	**83.6**	**47.2**
-single	98.5	97.5	67.5	44.3	96.7	93.5	52.7	30.7	96.4	93.6	38.2	17.9	95.3	85.6	27.7	16.7	99.2	96.9	37.5	27.6	97.1	88.9	16.8	13.4	94.9	79.7	6.9	4.9	92.3	67.7	2.3	1.9	96.3	87.9	31.2	19.7
-multi	99.7	96.2	87.5	60.9	99.3	93.3	86.1	57.3	99.3	89.5	85.7	47.0	92.5	86.8	75.1	**57.5**	**99.9**	97.9	64.6	26.6	**99.5**	91.7	38.1	17.5	98.1	83.3	25.1	12.3	89.4	71.8	18.4	12.3	97.2	88.8	60.1	36.4
-layer	99.9	95.7	91.1	63.5	99.9	92.9	65.2	49.2	**100.0**	97.3	66.1	35.3	99.5	85.7	56.0	19.3	99.5	98.8	73.5	45.7	98.7	96.1	42.7	23.7	96.1	91.9	34.7	18.7	88.9	**79.4**	21.6	12.4	97.8	92.2	56.4	33.5

Table 11: Detailed zero-shot scores on the NIAH benchmark for SWA hybridization methods for Qwen3-14B-Base under 4 SWA ρ ratios. The highest score under each configuration is highlighted in bold.

Method	Single-document QA				Multi-document QA				Summarization				Few-shot learning				Synthetic Tasks				Code Completion				Math Reasoning				Avg.			
	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875
Original	43.7				43.5				27.0				72.1				66.7				66.6				89.2				58.4			
Layer-level Heuristics																																
RAND	32.1	21.9	17.6	15.8	26.7	17.1	15.2	14.7	26.3	24.2	22.4	21.6	68.0	58.6	52.0	49.8	42.5	9.8	4.0	3.7	58.2	**56.4**	55.0	57.0	84.0	75.7	63.8	57.0	48.2	37.6	32.8	31.4
BME	25.1	5.3	15.8	14.6	23.2	3.5	10.8	11.7	18.8	11.6	21.4	21.7	47.9	18.5	46.4	50.4	3.5	2.3	3.7	3.8	41.2	25.4	50.0	53.9	68.9	45.0	55.2	55.3	32.6	15.9	29.1	30.2
INTR	20.3	5.3	16.6	18.4	19.0	4.6	14.0	14.1	25.4	11.1	21.7	21.5	65.4	35.3	53.8	53.3	27.8	2.0	2.8	3.0	61.2	48.5	**57.8**	53.1	84.6	68.0	49.7	49.4	43.4	25.0	30.9	30.4
Head-level Static Methods																																
DCAM	30.2	34.5	18.9	14.4	36.6	30.4	14.1	11.0	**26.9**	25.0	20.6	21.4	70.2	64.8	53.3	49.8	53.5	12.0	2.6	2.5	64.0	47.6	50.0	53.2	85.7	81.9	75.5	65.6	52.5	42.3	33.6	31.1
APL	42.1	36.6	15.8	14.8	38.8	**32.8**	9.2	10.4	26.7	23.8	17.3	21.0	**71.1**	**66.7**	45.5	48.6	56.5	21.0	4.8	3.3	51.9	42.5	45.0	54.3	85.1	82.0	72.5	53.3	53.2	43.6	30.0	29.4
QADA	35.6	26.7	20.8	15.5	29.7	21.3	15.4	12.5	26.4	23.8	20.2	19.5	67.9	59.0	49.0	44.9	18.8	3.7	2.7	3.0	47.6	42.0	43.8	46.6	81.1	78.1	66.6	54.6	43.9	36.4	31.2	28.1
RAZOR	42.2	30.3	18.9	17.8	42.0	31.3	15.6	13.1	26.2	24.2	20.4	21.6	69.0	64.8	55.7	54.0	53.4	35.2	**13.6**	5.2	48.7	50.9	54.3	**59.4**	84.0	82.5	72.8	59.7	52.2	45.6	35.9	33.0
FISHER	42.3	31.4	20.8	16.1	**42.5**	20.0	14.4	12.0	26.0	24.5	19.7	20.5	70.3	65.2	58.0	53.1	**61.0**	12.3	2.9	2.3	60.5	42.5	38.2	48.0	84.5	81.5	74.2	62.2	55.3	39.6	32.6	30.6
PROXY	40.1	30.2	15.9	15.6	39.1	28.6	12.4	10.2	25.9	23.9	19.4	17.4	69.9	63.0	48.5	53.0	42.8	4.7	2.6	3.2	47.5	43.4	45.5	40.9	83.5	84.3	79.4	63.8	49.8	39.7	32.0	29.2
Ours																																
BOSCH	**43.7**	**39.9**	**29.1**	**23.6**	41.8	30.0	**18.4**	**16.7**	26.1	**25.7**	**23.2**	**22.7**	69.5	66.5	**61.8**	**58.2**	57.8	28.2	10.0	**5.7**	**67.0**	53.0	43.0	52.8	**87.2**	**86.0**	**80.4**	**73.2**	**56.2**	**47.0**	**38.0**	**36.1**
-single	24.7	15.6	14.3	9.4	34.8	10.4	11.1	8.4	26.3	23.2	18.5	14.6	67.7	64.2	55.0	35.8	55.9	8.2	3.2	2.8	58.2	54.5	52.2	42.6	84.8	81.1	73.5	64.1	50.3	36.7	32.5	25.4
-multi	35.1	21.4	25.5	9.1	32.5	16.0	18.3	9.8	21.4	23.4	21.2	13.8	64.1	59.6	53.7	40.8	56.1	13.9	4.1	2.7	63.1	51.0	53.3	47.1	85.5	80.7	72.9	59.7	51.1	38.0	35.6	26.2
-layer	32.8	33.6	21.4	16.9	37.8	29.9	14.3	14.7	26.5	25.0	22.3	21.8	69.5	64.1	52.7	54.7	49.0	**38.9**	8.8	5.4	66.0	53.0	57.0	55.8	84.5	82.9	71.2	56.3	52.3	46.8	35.4	32.2

Table 12: Detailed zero-shot scores on the LongBench benchmark for SWA hybridization methods for Qwen3-14B-Base under 4 SWA ρ ratios. The highest score under each configuration is highlighted in bold.

Method	single																multikey																Avg.			
	4k				8k				16k				32k				4k				8k				16k				32k							
	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875				
Original	100.0				100.0				99.9				100.0				99.7				99.6				99.1				96.7				99.4			
Layer-level Heuristics																																				
RAND	100.0	31.0	30.8	30.6	99.8	22.2	21.9	18.0	99.9	20.0	19.8	10.8	97.3	9.7	9.8	2.9	87.7	22.0	19.9	19.5	61.9	8.6	7.9	7.9	51.6	4.2	4.3	3.7	35.5	1.3	1.4	1.2	79.6	16.0	15.0	10.6
BME	80.1	45.7	32.3	29.7	87.9	43.3	15.3	12.8	88.1	37.9	6.4	6.2	45.2	23.2	2.8	2.8	38.5	17.9	23.8	23.5	18.3	11.6	10.2	9.9	15.9	6.7	3.6	3.9	6.2	2.3	1.3	1.5	47.5	18.9	12.0	10.5
INTR	91.3	30.9	59.8	32.5	83.3	15.1	43.4	19.8	81.4	7.3	38.1	14.1	80.6	2.8	22.7	2.9	68.9	26.2	23.1	17.5	55.7	11.4	7.3	4.5	47.9	4.9	3.9	3.0	43.5	1.6	1.1	0.6	69.1	12.8	22.8	11.3
Head-level Static Methods																																				
DCAM	**100.0**	99.8	71.0	31.3	**100.0**	**99.9**	**90.7**	12.3	99.9	**100.0**	54.1	6.4	99.9	**99.9**	36.6	2.0	98.4	92.3	31.7	15.7	93.7	67.9	27.4	5.7	84.1	51.9	15.2	2.5	65.5	37.6	2.9	0.1	92.7	81.2	41.2	9.5
APL	99.9	92.9	32.1	31.8	99.9	89.6	14.4	12.8	98.1	74.3	7.3	6.9	88.9	50.6	2.8	2.5	99.0	48.7	16.7	16.2	87.8	22.4	6.3	6.2	82.4	13.3	2.9	3.0	56.5	9.3	0.8	0.6	89.1	50.1	10.4	10.0
PROXY	**100.0**	99.2	55.7	32.0	**100.0**	95.8	48.9	13.0	99.9	96.7	44.0	7.1	99.7	89.5	35.0	2.7	96.4	88.3	23.0	16.4	82.7	63.3	15.4	6.4	70.3	46.9	9.1	3.2	57.5	35.4	2.8	0.8	88.3	76.9	29.2	10.2
QADA	99.9	97.5	68.1	32.6	99.9	94.9	55.9	13.6	99.5	95.4	55.7	7.7	99.5	81.9	39.3	3.3	91.8	54.1	26.3	17.0	62.6	32.9	15.1	7.0	52.9	20.1	13.9	3.8	40.1	17.6	4.9	1.4	80.8	61.8	34.9	10.8
RAZOR	99.9	**99.9**	86.5	33.4	**100.0**	99.7	64.9	14.4	**100.0**	99.4	48.9	8.5	99.3	97.9	38.7	4.1	92.5	87.9	34.5	17.8	80.7	62.8	10.5	7.8	65.3	49.2	8.5	4.6	42.1	30.0	5.1	2.2	85.0	78.4	37.2	11.6
FISHER	**100.0**	99.7	72.3	32.9	**100.0**	99.6	68.7	13.9	**100.0**	99.1	70.3	8.0	99.9	99.4	43.1	3.6	**99.1**	88.5	**59.1**	17.3	**95.5**	67.2	**35.8**	7.3	84.3	56.7	**20.1**	4.1	63.3	41.5	7.7	1.7	92.8	81.5	47.1	11.1
BOSCH																																				
BOSCH	**100.0**	94.9	**90.1**	**62.6**	99.7	94.5	89.2	**57.0**	99.9	92.5	**74.6**	34.2	**100.0**	86.7	**55.6**	24.9	**99.7**	96.1	33.1	17.0	**98.1**	**88.3**	24.3	12.7	96.9	**76.9**	18.6	5.5	**85.7**	**60.6**	**16.1**	1.4	**97.5**	**86.3**	**50.2**	**26.9**
-single	97.4	91.4	62.6	41.3	92.8	82.8	57.0	31.9	93.3	76.3	34.2	26.4	87.2	60.9	24.9	17.5	86.3	59.4	17.0	**30.2**	66.4	47.0	12.7	13.0	53.1	32.1	5.5	5.4	33.7	16.3	1.4	**2.4**	76.3	58.3	26.9	21.0
-multi	99.5	97.3	57.3	49.0	99.2	95.2	56.1	36.1	97.9	91.4	42.9	36.9	95.0	84.0	39.4	22.3	96.9	82.7	28.4	22.0	87.7	57.3	14.9	**14.1**	70.0	42.2	7.2	**7.5**	40.1	29.3	5.9	1.8	85.8	72.4	31.5	23.7
-layer	99.3	98.3	82.4	49.3	98.2	96.9	66.0	39.9	97.8	96.7	60.3	**38.7**	95.1	87.7	41.4	**27.5**	98.0	**96.3**	34.1	16.0	98.0	82.7	20.1	7.5	**97.1**	67.6	12.2	4.8	85.3	54.1	6.3	1.3	96.1	85.0	40.3	23.1

Table 13: Detailed zero-shot scores on the NIAH benchmark for SWA hybridization methods for Qwen3-30B-A3B-Base under 4 SWA ρ ratios. The highest score under each configuration is highlighted in bold.

Method	Single-document QA				Multi-document QA				Summarization				Few-shot learning				Synthetic Tasks				Code Completion				Math Reasoning				Avg.			
	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875	0.25	0.5	0.75	0.875
Original	24.6				39.8				27.6				70.0				64.3				67.8				90.1				54.9			
Layer-level Heuristics																																
RAND	**14.3**	4.6	4.0	3.2	18.6	3.9	3.9	3.6	26.2	11.4	8.6	7.0	67.7	28.1	19.5	17.8	26.6	3.1	2.7	3.3	58.8	31.3	25.8	25.3	69.4	45.6	44.7	36.4	40.2	18.3	15.6	13.8
BME	9.5	4.2	4.2	3.7	5.9	3.9	3.8	3.7	21.9	10.1	7.3	8.4	55.1	28.9	25.0	20.9	3.1	2.7	3.2	3.0	32.0	23.6	17.4	30.6	70.1	43.1	25.5	27.3	28.2	16.7	12.3	14.0
INTR	12.6	5.2	4.7	3.3	17.5	3.9	4.3	4.0	25.7	15.7	9.8	7.1	68.7	36.0	24.9	24.4	23.9	2.8	3.0	3.2	65.5	52.7	34.0	26.6	**82.5**	60.2	24.0	23.0	42.3	25.2	14.9	13.1
Head-level Static Methods																																
DCAM	13.2	11.8	**9.7**	1.3	23.6	15.2	**9.3**	2.1	**26.6**	23.9	16.0	5.5	68.5	58.8	43.5	19.5	27.1	3.0	3.1	2.2	49.8	44.5	34.0	21.7	77.0	60.3	31.9	22.0	40.8	31.1	21.1	10.6
APL	8.6	8.0	5.6	2.3	9.2	6.2	3.5	3.1	25.2	19.6	7.2	6.5	66.1	52.0	15.0	20.5	36.6	2.7	3.3	3.2	**68.0**	48.2	22.4	22.7	78.8	61.0	35.9	23.0	41.8	28.3	13.3	11.6
PROXY	9.4	8.5	5.9	2.7	12.2	6.9	4.7	3.5	25.8	21.0	10.8	6.9	65.5	52.8	28.9	20.9	26.5	4.4	3.7	3.6	53.1	34.5	29.1	23.1	76.3	60.9	37.3	23.4	38.4	27.0	17.2	12.0
QADA	12.0	9.6	5.6	3.6	21.4	8.6	5.9	4.4	25.3	20.0	15.2	7.8	67.4	59.6	41.2	21.8	21.3	2.9	2.2	4.5	38.5	31.1	28.1	24.0	60.0	56.9	44.4	24.3	35.1	27.0	20.4	12.9
RAZOR	12.8	6.7	5.9	5.0	16.8	5.0	5.7	5.8	26.0	22.9	15.6	9.2	65.9	53.9	37.0	23.2	18.7	4.6	2.3	5.9	41.2	**57.5**	28.9	25.4	59.9	63.0	36.5	25.7	34.5	30.5	18.8	14.3
FISHER	14.1	**13.7**	7.7	4.5	23.6	**16.4**	7.1	5.3	25.9	23.0	16.6	8.7	**69.7**	58.2	43.2	22.7	38.9	4.8	3.2	**5.4**	60.1	38.8	36.1	24.9	77.2	67.8	32.1	25.2	44.2	31.8	20.9	13.8
BOSCH																																
BOSCH	12.3	12.8	7.1	**5.5**	**26.1**	**16.4**	7.5	4.3	26.5	**24.0**	**16.9**	**11.0**	69.2	**64.8**	**45.7**	28.3	46.3	**7.1**	3.0	2.8	61.6	49.0	32.8	32.4	81.0	**77.6**	**60.7**	**51.0**	**46.2**	**36.0**	**24.8**	**19.3**
-single	8.5	6.8	4.2	3.5	8.9	4.5	4.4	3.7	25.7	17.5	8.8	7.2	59.2	46.6	33.4	26.0	20.6	3.2	**4.3**	2.9	57.2	41.2	18.9	20.0	57.7	71.9	51.0	35.3	34.0	27.4	17.9	14.1
-multi	9.2	7.0	5.5	4.2	12.9	6.6	4.3	**4.7**	26.1	22.0	11.0	8.5	66.5	54.0	28.3	**28.6**	37.4	3.0	2.8	3.0	61.9	46.8	32.4	27.2	71.6	57.9	37.2	26.9	40.8	28.2	17.4	14.7
-layer	7.8	7.4	6.2	4.6	9.9	5.7	4.4	3.8	25.2	21.9	11.4	10.5	59.8	53.7	35.7	22.6	**47.0**	3.2	3.7	2.9	63.6	51.7	**38.8**	**34.3**	66.0	72.2	40.4	26.7	39.9	30.8	20.1	15.1

Table 14: Detailed zero-shot scores on the LongBench benchmark for SWA hybridization methods for Qwen3-30B-A3B-Base under 4 SWA ρ ratios. The highest score under each configuration is highlighted in bold.

Method	single																multikey																Avg.			
	4k				8k				16k				32k				4k				8k				16k				32k							
	0.7B	1.3B	2.0B	2.5B	0.7B	1.3B	2.0B	2.5B	0.7B	1.3B	2.0B	2.5B	0.7B	1.3B	2.0B	2.5B	0.7B	1.3B	2.0B	2.5B	0.7B	1.3B	2.0B	2.5B	0.7B	1.3B	2.0B	2.5B	0.7B	1.3B	2.0B	2.5B	0.7B	1.3B	2.0B	2.5B
Qwen3-1.7B-Base																																				
Original	100.0				99.9				100.0				100.0				98.9				94.9				84.1				63.9				92.7			
INTR	95.9	96.4	96.5	95.7	91.6	93.4	95.1	93.6	72.6	75.5	78.7	76.7	52.2	57.3	59.4	56.4	80.3	87.3	88.4	88.9	64.1	74.9	78.9	79.1	36.4	49.5	56.9	57.3	17.9	24.7	30.1	31.7	63.9	69.9	73.0	72.4
B-layer	**99.7**	99.0	98.7	98.4	**99.7**	96.1	94.1	94.5	99.4	93.1	89.9	89.8	98.9	90.8	84.2	85.5	98.2	99.2	99.2	99.3	89.5	93.1	93.0	93.5	79.4	85.9	85.5	86.6	70.2	72.5	72.9	72.6	91.9	91.2	89.7	90.0
FISHER	99.5	**99.9**	**99.7**	**99.9**	98.9	**98.3**	**99.6**	**98.6**	97.3	97.3	99.5	97.9	88.1	89.1	**97.6**	89.5	**99.3**	99.3	98.1	99.6	**97.3**	97.3	95.9	97.7	90.5	90.2	87.7	92.0	**75.8**	**77.3**	74.1	77.9	93.3	93.6	94.0	94.1
BOSCH	98.4	99.2	99.5	99.4	97.3	95.1	95.8	97.1	**99.9**	**99.8**	**100.0**	**100.0**	**99.2**	**97.9**	96.3	**94.7**	98.6	**99.5**	**99.7**	**99.7**	95.7	**97.5**	**98.2**	**98.7**	**91.0**	**92.7**	**94.1**	**94.1**	68.0	74.2	**76.0**	**78.1**	**93.5**	**94.5**	**94.9**	**95.2**
Qwen3-8B-Base																																				
Original	100.0				100.0				100.0				99.9				99.5				99.6				99.1				94.5				99.1			
INTR	99.6	99.9	99.3	99.4	99.4	99.5	99.1	99.5	97.9	98.3	95.2	97.2	96.1	96.8	93.0	95.7	94.4	96.5	98.4	98.2	88.3	89.5	91.9	92.3	66.3	68.9	76.8	79.2	49.1	52.9	61.8	65.3	86.4	87.8	89.4	90.8
B-layer	**100.0**	**100.0**	**100.0**	99.9	**99.9**	**100.0**	99.5	99.6	**99.9**	97.8	94.5	95.7	99.3	90.6	85.3	87.0	93.3	98.3	99.2	98.9	85.9	94.8	96.5	96.2	66.3	81.0	85.5	86.1	45.9	60.4	67.3	69.1	86.3	90.4	91.0	91.6
FISHER	99.7	99.9	99.9	99.9	99.5	100.0	99.9	**99.9**	98.6	99.7	99.7	99.8	97.9	99.5	99.7	99.3	99.6	**99.8**	**99.9**	**99.9**	98.7	99.4	99.6	99.5	96.3	98.4	**99.1**	**99.5**	84.5	90.5	92.4	93.3	96.9	98.5	98.8	98.9
BOSCH	100.0	100.0	100.0	**100.0**	99.9	99.9	**100.0**	99.9	99.9	**100.0**	**100.0**	**100.0**	**99.5**	99.8	**99.8**	**99.9**	**99.9**	**99.8**	99.9	99.9	**98.8**	**99.7**	**99.7**	**99.7**	**97.8**	**98.9**	99.1	99.1	**90.0**	**91.3**	**93.8**	**94.0**	**98.2**	**98.6**	**99.0**	**99.1**
Qwen3-14B-Base																																				
Original	100.0				99.9				100.0				100.0				99.8				99.7				99.3				98.2				99.6			
INTR	99.8	97.1	99.7	99.3	98.9	93.9	98.8	98.7	91.5	86.4	92.1	89.6	79.4	64.6	74.1	70.8	98.3	98.0	99.1	99.6	92.3	89.8	92.5	94.6	65.9	67.7	68.1	73.1	38.3	37.7	44.7	49.1	83.0	79.4	83.6	84.4
B-layer	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	99.6	**100.0**	**100.0**	96.9	92.8	97.3	97.7	97.1	99.3	98.9	99.6	92.3	96.1	95.3	98.1	79.5	80.3	84.9	86.7	59.7	60.0	61.7	66.5	90.7	91.0	92.3	93.6
FISHER	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	99.8	**100.0**	99.9	99.9	99.9	99.9	**100.0**	**100.0**	99.9	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**99.9**	**100.0**	**99.9**	**100.0**	**99.7**	**99.9**	**99.9**	99.8	97.3	97.7	98.3	**99.0**	99.6	99.7	99.7	**99.8**
BOSCH	**100.0**	**100.0**	**100.0**	**100.0**	99.9	**100.0**	**100.0**	**100.0**	99.9	**100.0**	**100.0**	**100.0**	99.7	99.9	**100.0**	**100.0**	**100.0**	99.9	**100.0**	**100.0**	**99.9**	99.7	**99.9**	99.9	**99.7**	99.8	99.8	**99.9**	**98.5**	**98.7**	**98.8**	98.7	**99.7**	**99.8**	**99.8**	**99.8**
Qwen3-30B-A3B-Base																																				
Original	100.0				100.0				99.9				100.0				99.7				99.6				99.1				96.7				99.4			
INTR	99.6	**100.0**	99.9	99.9	97.8	99.4	99.0	99.4	95.1	98.7	98.4	97.7	88.1	91.5	91.6	95.5	99.7	99.7	99.8	99.9	96.0	98.1	98.0	97.8	86.5	85.5	87.4	86.5	71.5	72.1	73.1	72.9	91.8	93.1	93.4	93.7
B-layer	99.1	99.9	**100.0**	**100.0**	99.3	99.9	**100.0**	**100.0**	97.3	99.8	**99.9**	**100.0**	88.7	98.0	98.2	98.9	99.9	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	98.7	99.4	**99.5**	**99.7**	92.5	96.3	**97.5**	**97.5**	70.4	77.7	78.3	77.5	93.2	96.4
FISHER	**100.0**	**100.0**	**100.0**	**100.0**	99.9	99.9	99.9	**100.0**	99.4	99.7	99.7	99.6	95.6	98.3	98.4	98.8	**100.0**	**100.0**	**100.0**	**100.0**	99.9	99.3	98.3	**99.5**	99.5	97.1	96.4	97.7	96.3	84.5	83.7	87.3	93.3	97.0	97.0	97.8	98.4
BOSCH	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	99.8	**100.0**	**100.0**	**100.0**	**99.9**	**100.0**	**99.6**	**98.5**	**99.2**	**99.6**	**100.0**	99.9	99.8	**100.0**	99.6	99.8	**99.5**	**99.6**	97.9	98.0	**98.3**	**98.7**	88.3	91.6	**91.6**	**94.2**	**98.2**	**98.5**	**98.5**	**99.0**

Table 15: Detailed continual pretraining performances on the NIAH benchmark for 4 SWA hybridization methods for 4 Qwen3 models (1.7B, 8B, 14B, 30B) under SWA ratio $\rho = 0.75$. The highest score under each configuration is highlighted in bold.

Method	Single-document QA				Multi-document QA				Summarization				Few-shot learning				Synthetic Tasks				Code Completion				Math Reasoning				Avg.			
	0.7B	1.3B	2.0B	2.5B	0.7B	1.3B	2.0B	2.5B	0.7B	1.3B	2.0B	2.5B	0.7B	1.3B	2.0B	2.5B	0.7B	1.3B	2.0B	2.5B	0.7B	1.3B	2.0B	2.5B	0.7B	1.3B	2.0B	2.5B	0.7B	1.3B	2.0B	2.5B
Qwen3-1.7B-Base																																
Original	39.8				30.9				26.0				67.4				32.7				48.2				72.6				45.4			
INTR	25.9	28.1	28.4	29.4	22.6	23.8	25.3	25.2	20.6	21.1	21.3	21.5	55.6	55.8	57.0	57.9	3.8	4.7	4.7	4.7	50.8	47.3	46.8	45.8	58.7	59.9	60.6	61.1	34.0	34.4	34.9	35.1
B-layer	**35.4**	36.5	37.3	37.2	25.4	27.4	**29.4**	**30.3**	**21.6**	22.4	22.8	22.6	60.8	62.4	62.6	63.0	6.0	10.0	**13.0**	**13.2**	48.2	46.9	46.1	44.2	57.8	60.3	59.8	61.1	36.9	38.4	38.7	38.8
FISHER	35.3	**36.8**	**37.9**	38.1	24.6	27.1	28.8	28.3	21.3	22.7	23.5	23.5	**62.9**	**65.2**	**65.4**	**65.4**	7.2	8.3	8.8	9.2	44.3	42.4	43.7	44.2	**59.5**	**61.3**	60.4	60.7	36.4	37.7	38.4	38.5
S1&3	34.5	36.6	**37.9**	**38.2**	**26.2**	**28.8**	28.9	28.9	21.2	**23.8**	**24.0**	**23.9**	61.2	62.3	63.1	63.0	**8.3**	**13.5**	10.2	10.5	**51.6**	**47.4**	**47.3**	**46.4**	59.2	60.0	**62.2**	**62.2**	**37.0**	**38.5**	**39.1**	**39.0**
Qwen3-8B-Base																																
Original	45.5				41.9				26.8				70.3				63.0				66.6				85.8				57.1			
INTR	32.4	33.0	33.7	33.9	32.0	29.9	30.7	31.9	19.6	18.4	19.2	19.8	62.5	63.9	64.0	64.7	6.5	6.3	7.5	7.8	56.5	53.0	56.2	54.4	81.4	82.0	**83.5**	82.4	41.6	40.9	42.1	42.1
B-layer	39.9	40.1	40.8	40.8	33.9	34.5	34.3	35.5	24.6	22.4	23.4	23.9	63.1	63.9	64.3	64.6	15.3	17.0	18.7	19.7	53.8	**55.4**	**58.8**	**56.8**	80.5	82.3	80.8	81.1	44.4	45.1	45.9	46.1
FISHER	44.0	45.3	44.9	44.9	**38.4**	38.2	36.2	37.3	24.7	22.7	22.8	22.7	**67.3**	**67.8**	**68.5**	**68.6**	17.2	19.8	19.7	19.8	52.1	55.0	56.9	55.9	**81.6**	82.6	81.5	82.4	46.5	47.4	47.2	47.4
BOSCH	**45.5**	**46.1**	**46.3**	**46.0**	36.2	**38.7**	**37.6**	**38.3**	**25.6**	**24.6**	**24.7**	**24.5**	64.9	65.3	65.7	65.4	**34.5**	**37.5**	**41.3**	**41.0**	**58.7**	52.8	54.8	54.9	79.9	**84.0**	83.4	**83.5**	**49.3**	**49.9**	**50.6**	**50.5**
Qwen3-14B-Base																																
Original	43.7				43.5				27.0				72.1				66.7				66.6				89.2				58.4			
INTR	28.9	32.1	32.0	33.2	28.8	28.5	28.5	29.7	13.6	16.5	16.6	15.0	59.3	62.8	63.4	62.1	11.0	9.3	11.3	10.7	58.0	46.5	51.4	52.8	0.2	2.5	4.8	18.2	28.5	28.3	29.7	31.7
B-layer	36.9	38.9	41.4	**44.5**	31.5	34.1	35.4	35.7	10.0	13.9	15.3	16.7	66.9	67.2	**71.2**	**69.0**	16.0	16.3	41.5	42.3	60.5	57.1	**61.0**	52.6	0.7	9.9	10.5	27.0	31.8	33.9	39.5	41.1
FISHER	**42.8**	**42.8**	37.5	39.8	**35.0**	**35.5**	33.0	33.2	16.8	17.3	13.9	11.9	**68.9**	**69.0**	66.5	66.3	30.3	34.3	21.7	23.5	51.5	50.9	56.9	**60.0**	6.7	8.3	47.4	60.8	36.0	36.9	39.6	42.2
BOSCH	40.4	42.3	**44.0**	42.6	32.5	33.6	**35.8**	**36.2**	16.0	14.5	16.8	17.6	67.7	68.5	68.3	68.8	**48.2**	**48.2**	**53.8**	**55.2**	**66.0**	**57.6**	58.6	56.6	1.1	**52.5**	**81.9**	**83.5**	**38.8**	**45.3**	**51.3**	**51.5**
Qwen3-30B-A3B-Base																																
Original	24.6				39.8				27.6				70.0				64.3				67.8				90.1				54.9			
INTR	**19.8**	**26.5**	**28.4**	17.9	17.5	**18.7**	19.1	19.2	13.7	13.6	13.0	17.2	47.2	**52.2**	51.5	**52.4**	14.6	15.1	18.6	25.6	54.0	**52.7**	52.3	**52.8**	40.1	45.5	**62.8**	61.1	29.6	**32.0**	35.1	35.2
B-layer	10.3	12.7	17.9	12.7	**18.4**	17.7	**19.2**	17.9	15.2	17.5	**17.2**	**17.4**	**51.6**	49.3	**52.4**	48.3	**20.9**	25.5	**25.6**	29.7	**54.7**	50.3	52.8	48.0	41.1	46.7	61.1	**79.0**	30.3	31.4	**35.2**	36.1
FISHER	11.3	10.1	11.1	11.8	17.1	16.4	18.3	18.2	15.4	12.3	13.4	14.8	48.7	50.0	49.9	49.9	18.6	**25.8**	21.8	**30.7**	49.8	50.5	51.1	51.2	44.8	49.7	54.3	75.5	29.4	30.7	31.4	36.0
BOSCH	9.2	11.2	27.7	**22.1**	16.7	17.1	18.9	**21.5**	**17.8**	**19.0**	14.3	16.4	48.1	49.9	51.9	50.6	17.9	21.7	15.4	25.7	54.5	45.2	**53.2**	**52.8**	**49.1**	**54.7**	50.9	71.6	**30.5**	31.3	33.2	**37.2**

Table 16: Detailed continual pretraining performances on the LongBench benchmark for 4 SWA hybridization methods for 4 Qwen3 models (1.7B, 8B, 14B, 30B) under SWA ratio $\rho = 0.75$. The highest score under each configuration is highlighted in bold.

Method	single																multikey																Avg.			
	4k				8k				16k				32k				4k				8k				16k				32k							
	256	512	2k	4k	256	512	2k	4k	256	512	2k	4k	256	512	2k	4k	256	512	2k	4k	256	512	2k	4k	256	512	2k	4k	256	512	2k	4k	256	512	2k	4k
Qwen3-1.7B-Base																																				
Original	100.0				99.9				100.0				100.0				98.9				94.9				84.1				63.9				92.7			
INTR	5.1	13.8	55.6	**100.0**	3.2	7.3	30.9	55.5	2.4	2.7	14.2	28.9	2.1	2.3	6.8	13.1	4.1	9.5	40.7	**98.9**	2.6	6.1	16.3	27.6	1.6	2.3	7.9	13.5	0.9	1.2	3.9	5.5	2.8	5.7	22.0	42.9
B-layer	**82.3**	**83.9**	89.3	**100.0**	**80.1**	**81.1**	**85.7**	89.2	**73.2**	**76.2**	**81.8**	**86.5**	**64.2**	59.6	59.3	66.7	27.1	27.7	57.7	**98.9**	18.1	22.2	26.0	36.3	20.3	18.9	20.8	25.3	14.2	13.4	15.8	17.8	**47.4**	47.9	54.5	65.1
FISHER	66.7	80.5	**95.2**	**100.0**	66.5	75.7	83.4	**93.3**	53.3	69.2	75.2	82.4	27.0	38.1	68.9	77.8	21.2	30.4	60.2	**98.9**	17.1	24.3	32.2	33.7	15.9	18.5	24.4	25.1	3.5	6.6	18.7	21.7	33.9	42.9	57.3	66.6
BOSCH	70.1	78.4	84.9	**100.0**	67.4	77.2	81.4	84.5	52.4	63.3	80.0	82.8	47.5	**60.5**	**74.9**	**81.1**	**41.2**	**47.4**	**64.0**	**98.9**	**30.2**	**39.3**	**41.9**	**36.7**	**20.8**	**27.3**	**34.5**	**36.1**	**18.7**	**24.8**	**32.4**	**31.1**	43.5	**52.3**	**61.8**	**68.9**
Qwen3-8B-Base																																				
Original	100.0				100.0				100.0				99.9				99.5				99.6				99.1				94.5				99.1			
INTR	5.1	13.9	32.3	**100.0**	3.5	7.4	15.6	55.8	2.4	2.6	7.6	28.9	2.1	2.3	2.9	13.2	4.6	12.2	23.9	**99.5**	2.7	6.7	12.5	30.5	1.7	2.3	4.7	18.1	1.0	1.2	1.7	7.4	2.9	6.1	12.7	44.2
B-layer	68.2	62.4	74.0	**100.0**	49.0	51.5	76.0	74.6	34.7	51.3	64.7	52.4	31.6	37.6	53.3	41.4	11.9	17.1	63.1	**99.5**	6.1	10.4	21.1	30.1	2.1	8.2	12.6	19.4	1.7	2.6	8.5	7.5	25.7	30.1	46.7	53.1
FISHER	86.2	88.7	94.9	**100.0**	81.3	81.3	94.3	94.4	**87.3**	65.2	91.9	89.5	53.7	64.5	78.1	**92.4**	37.7	57.7	83.4	**99.5**	27.2	35.9	61.3	66.4	**24.2**	28.1	45.6	51.7	12.1	18.7	27.7	**37.6**	51.2	55.0	72.2	78.9
BOSCH	**97.7**	**99.1**	**98.7**	**100.0**	**91.3**	**92.3**	**98.1**	**97.5**	82.9	**93.1**	**97.8**	**97.3**	**63.4**	**79.8**	**90.9**	90.5	**46.2**	**65.1**	**96.1**	**99.5**	**28.1**	**39.4**	**65.0**	**82.3**	22.9	**29.3**	**48.1**	**58.5**	**16.5**	**22.7**	**31.1**	36.3	**56.1**	**65.1**	**78.2**	**82.7**
Qwen3-14B-Base																																				
Original	100.0				99.9				100.0				100.0				99.8				99.7				99.3				98.2				99.6			
INTR	5.0	13.8	56.6	**100.0**	3.5	7.3	31.0	55.9	2.4	2.6	14.5	28.9	2.1	2.3	6.9	13.3	3.1	8.9	56.1	99.7	2.3	5.5	16.3	26.1	1.4	1.8	7.3	16.7	0.8	0.9	4.3	6.3	2.6	5.4	24.1	43.4
B-layer	33.9	51.3	86.5	98.3	36.9	44.0	91.8	78.5	32.8	33.1	53.3	58.0	27.7	30.3	47.8	37.9	5.9	15.9	85.2	98.3	4.1	8.2	45.4	49.9	2.3	2.7	30.2	31.9	1.2	1.6	23.3	15.1	18.1	23.4	57.9	60.4
FISHER	90.9	**94.8**	**97.1**	**100.0**	88.7	93.5	**97.2**	**97.7**	**77.9**	**91.9**	92.2	95.1	54.0	**78.7**	**91.1**	**91.1**	38.1	57.1	96.1	**99.8**	28.1	39.5	59.1	81.5	18.8	29.0	44.9	52.9	10.8	23.0	35.3	40.9	50.9	63.5	76.6	82.4
BOSCH	**91.2**	91.7	95.7	**100.0**	**91.1**	**94.0**	96.6	97.2	75.1	87.0	**97.5**	**97.4**	**56.5**	68.1	83.8	91.0	**70.1**	**84.0**	**98.1**	99.7	**54.7**	**66.3**	**85.1**	**94.8**	**47.5**	**59.5**	**78.4**	**85.8**	**29.1**	**45.6**	**66.5**	**74.5**	**64.4**	**74.5**	**87.7**	**92.5**
Qwen3-30B-A3B-Base																																				
Original	24.6				39.8				27.6				70.0				64.3				67.8				90.1				54.9							
INTR	48.9	49.7	66.4	**100.0**	40.1	44.9	44.1	75.9	27.1	29.3	39.5	40.0	11.2	13.6	12.7	17.3	10.2	14.5	20.3	**99.7**	5.5	9.0	11.3	20.0	1.3	2.6	7.5	9.0	2.9	3.2	1.9	2.8	18.4	20.9	25.5	45.6
B-layer	58.3	69.7	83.1	**100.0**	44.7	52.4	77.1	**95.0**	38.5	42.6	71.7	83.3	37.1	38.8	**51.1**	45.6	14.7	22.0	28.5	**99.7**	9.5	13.5	26.5	37.1	4.3	7.3	17.0	24.1	4.2	5.2	**10.1**	12.9	26.4	31.4	45.6	62.2
FISHER	78.7	58.3	**92.7**	**100.0**	70.3	72.0	**90.3**	93.0	46.6	61.3	**89.2**	**89.1**	33.9	35.7	39.3	67.1	**30.7**	**44.3**	34.7	**99.7**	**18.7**	**28.5**	25.9	27.1	9.4	**15.1**	20.2	17.7	5.4	6.2	8.6	12.4	36.7	40.2	50.1	63.3
BOSCH	**84.4**	**89.8**	91.9	**100.0**	**77.2**	**84.5**	81.8	88.3	**78.1**	**79.0**	82.9	84.3	**53.9**	**59.4**	45.5	**71.3**	24.9	29.8	**80.1**	99.6	15.7	21.9	**50.1**	**63.5**	**17.3**	14.9	**23.0**	**33.4**	**10.0**	**12.0**	7.3	**16.1**	**45.2**	**48.9**	**57.8**	**69.6**

Table 17: Detailed zero-shot performances on the NIAH benchmark for 4 SWA hybridization methods for 4 Qwen3 models (1.7B, 8B, 14B, 30B) under SWA ratio $\rho = 0.75$ when SWA windows size is $\{256, 512, 2048, 4096\}$. The highest score under each configuration is highlighted in bold.

Method	Single-document QA				Multi-document QA				Summarization				Few-shot learning				Synthetic Tasks				Code Completion				Math Reasoning				Avg.			
	256	512	2k	4k	256	512	2k	4k	256	512	2k	4k	256	512	2k	4k	256	512	2k	4k	256	512	2k	4k	256	512	2k	4k	256	512	2k	4k
Qwen3-1.7B-Base																																
Original	39.8				30.9				26.0				67.4				32.7				48.2				72.6				45.4			
INTR	13.2	13.9	16.8	25.1	**11.4**	**11.9**	12.6	13.3	11.8	14.6	18.3	20.0	19.2	27.9	47.1	55.8	**3.1**	3.5	**3.8**	4.9	24.6	23.9	30.4	39.0	3.9	4.6	71.0	71.0	12.5	14.3	28.6	32.7
B-layer	13.3	**15.8**	21.6	27.9	8.9	10.9	13.7	16.0	14.0	**18.0**	**21.7**	**23.1**	29.2	34.6	49.5	55.3	3.0	**3.7**	3.6	**5.8**	31.2	32.2	37.2	43.6	**15.2**	**19.6**	71.0	71.0	16.4	**19.3**	31.2	34.7
FISHER	10.9	12.8	19.5	26.7	8.9	9.5	12.6	15.3	10.0	12.2	17.5	19.9	**32.4**	**40.8**	50.9	55.0	1.8	2.6	3.0	5.2	18.5	17.8	26.3	34.8	6.0	7.5	71.3	71.3	12.6	14.7	28.7	32.6
BOSCH	**14.8**	15.7	**23.0**	**29.4**	10.7	10.6	**16.1**	**18.4**	**15.0**	17.0	21.6	22.8	29.0	39.2	**54.8**	**58.9**	**3.1**	3.4	3.7	5.7	**32.5**	**33.3**	**41.4**	**43.9**	10.6	15.6	**71.5**	**71.5**	**16.5**	**19.3**	**33.2**	**35.8**
Qwen3-8B-Base																																
Original	45.5				41.9				26.8				70.3				63.0				66.6				85.8				57.1			
INTR	13.9	16.4	21.8	31.5	11.0	12.7	16.2	18.8	14.2	**18.9**	21.3	23.3	30.8	**41.9**	51.4	56.3	**3.6**	**4.6**	2.8	7.3	20.5	25.4	46.4	54.7	**49.5**	**57.9**	83.3	83.3	20.5	25.4	34.7	39.3
B-layer	5.1	6.1	14.6	28.8	6.3	6.7	7.6	9.9	15.8	17.3	21.3	23.4	30.0	37.2	55.3	58.6	2.9	2.6	3.2	12.5	**45.3**	**50.4**	**57.4**	**61.8**	35.9	48.1	**83.5**	**83.5**	20.2	24.1	34.7	39.8
FISHER	6.2	9.4	20.9	33.0	3.3	3.6	8.8	13.1	6.4	9.3	18.6	22.2	**35.0**	41.6	55.8	60.0	3.0	2.7	3.6	12.0	22.4	26.0	38.2	51.9	4.2	6.0	83.2	83.2	11.5	14.1	32.7	39.3
BOSCH	**14.6**	**16.7**	**27.0**	**35.0**	**14.0**	**14.4**	**18.9**	**23.2**	**16.7**	18.1	**23.5**	**24.9**	27.2	33.5	**56.3**	**63.1**	2.7	2.4	**6.4**	**15.7**	40.4	42.1	42.5	54.1	46.0	54.4	83.3	83.3	**23.1**	**25.9**	**36.8**	**42.8**
Qwen3-14B-Base																																
Original	43.7				43.5				27.0				72.1				66.7				66.6				89.2				58.4			
INTR	15.9	14.5	18.8	27.9	11.5	12.3	16.5	16.4	17.6	20.0	23.5	25.1	34.0	43.1	58.6	64.3	2.8	3.3	4.2	13.6	**50.9**	**53.1**	57.3	61.9	49.6	61.0	84.8	84.8	26.0	29.6	37.7	42.0
B-layer	**19.6**	10.0	14.3	25.8	**13.7**	8.3	13.8	18.9	14.8	17.4	23.0	**25.5**	**46.9**	53.1	**64.1**	**66.8**	3.0	3.8	7.3	24.6	33.0	47.9	**57.9**	**62.8**	41.8	54.1	84.8	84.8	24.7	27.8	37.9	44.2
FISHER	18.9	18.9	22.3	30.8	**13.7**	13.9	16.4	**20.6**	**17.8**	17.0	22.4	24.5	34.2	**53.6**	61.7	64.3	3.3	2.7	3.8	10.2	49.0	35.8	46.5	54.8	40.0	49.7	85.2	85.2	25.3	27.4	36.9	41.5
BOSCH	8.9	**19.0**	**25.4**	**32.9**	7.5	**14.8**	**17.6**	20.5	14.5	**20.5**	**24.1**	**25.5**	41.6	44.9	59.3	65.2	**4.4**	**4.6**	**17.6**	**32.9**	44.8	51.6	55.9	60.4	**64.3**	**66.0**	**85.4**	**85.4**	**26.6**	**31.6**	**40.7**	**46.1**
Qwen3-30B-A3B-Base																																
Original	24.6				39.8				27.6				70.0				64.3				67.8				90.1				54.9			
INTR	4.0	4.0	4.7	6.2	4.5	4.4	4.1	4.7	8.1	9.1	11.3	14.5	17.5	21.7	28.6	34.7	3.3	2.9	**3.1**	**6.2**	**33.8**	32.3	37.1	49.1	1.1	5.6	83.5	83.5	10.3	11.4	24.6	28.4
B-layer	4.9	5.7	7.0	8.5	4.3	4.5	4.1	4.7	8.2	10.1	14.5	17.5	24.4	29.4	41.2	49.1	**3.6**	**3.5**	**3.1**	5.8	33.5	**34.9**	**42.9**	**53.0**	2.4	7.1	83.5	83.5	11.6	13.6	28.0	31.7
FISHER	5.7	**6.8**	6.0	7.3	**6.6**	**7.0**	6.1	6.0	10.8	13.4	19.8	21.6	29.8	38.5	**49.0**	51.6	**3.6**	3.0	2.2	5.4	28.1	29.1	35.9	48.0	24.6	31.5	83.2	83.2	15.6	18.5	28.9	31.9
BOSCH	**5.9**	6.3	**7.3**	**8.8**	6.4	6.5	**7.4**	**6.9**	**12.9**	**14.8**	**20.2**	**23.3**	**37.7**	**42.7**	48.2	**55.0**	2.5	2.9	**3.1**	5.6	28.2	29.0	38.6	49.1	**26.6**	**37.0**	**83.9**	**83.9**	**17.2**	**19.9**	**29.8**	**33.2**

Table 18: Detailed zero-shot performances on the LongBench benchmark for 4 SWA hybridization methods for 4 Qwen3 models (1.7B, 8B, 14B, 30B) under SWA ratio $\rho = 0.75$ when SWA windows size is $\{256, 512, 2048, 4096\}$. The highest score under each configuration is highlighted in bold.

Method	single				multikey				Avg.
	4k	8k	16k	32k	4k	8k	16k	32k	
Qwen3-1.7B-Base									
$0.25 \rightarrow 0.5$	81.6	80.4	71.0	70.2	51.9	41.5	31.3	29.1	57.1
$0.5 \rightarrow 0.75$	69.8	61.0	52.9	45.3	53.6	44.5	29.7	20.0	47.1
$0.75 \rightarrow 0.875$	44.8	31.5	13.1	10.7	22.8	18.7	8.0	8.0	19.7
Qwen3-8B-Base									
$0.25 \rightarrow 0.5$	96.9	94.9	93.8	80.9	91.2	77.4	65.3	35.7	79.5
$0.5 \rightarrow 0.75$	93.3	91.1	87.0	70.6	80.3	59.8	43.3	23.8	68.7
$0.75 \rightarrow 0.875$	44.6	31.7	21.4	13.0	31.0	14.9	6.6	2.0	20.7
Qwen3-14B-Base									
$0.25 \rightarrow 0.5$	99.9	97.6	95.4	83.9	97.9	91.9	82.4	53.8	87.9
$0.5 \rightarrow 0.75$	92.9	79.2	93.9	65.4	92.9	68.0	59.0	40.1	73.9
$0.75 \rightarrow 0.875$	68.8	49.9	52.4	27.3	59.6	32.6	25.4	12.5	41.1
Qwen3-30B-A3B-Base									
$0.25 \rightarrow 0.5$	94.9	91.6	76.7	59.3	89.2	64.0	36.8	24.9	67.2
$0.5 \rightarrow 0.75$	69.1	61.0	52.2	30.7	41.8	22.8	12.5	4.0	36.8
$0.75 \rightarrow 0.875$	35.9	26.6	21.5	9.2	20.8	7.5	3.4	0.8	15.7

Table 19: Zero-shot scores on the NIAH benchmark for BOSCH when using $\mathcal{A}'_{\rho_l}$ across 4 Qwen3 models under 3 SWA ratios $\rho \in \{0.5, 0.75, 0.875\}$. The left side of $\rightarrow$ corresponds to ρ_s, while the right side corresponds to ρ_l. $\mathcal{A}'_{\rho_l}$ indicates results obtained after replacing a randomly selected subset of heads at a given configuration (e.g., $\rho_l = 0.75$) with heads that did not appear in a smaller configuration (e.g., $\rho_s = 0.5$). Scores are averaged over three runs with different seeds.

Method	Tasks							Avg.
	Single-document QA	Multi-document QA	Summarization	Few-shot learning	Synthetic Tasks	Code Completion	Math Reasoning	
Qwen3-1.7B-Base								
$0.25 \rightarrow 0.5$	15.0	8.9	18.3	48.3	3.8	47.4	49.8	27.4
$0.5 \rightarrow 0.75$	11.6	9.1	17.1	43.4	2.9	38.4	38.8	23.0
$0.75 \rightarrow 0.875$	4.8	3.2	12.9	39.9	2.2	24.0	31.6	17.0
Qwen3-8B-Base								
$0.25 \rightarrow 0.5$	22.2	16.1	22.4	58.4	10.2	42.6	78.1	34.3
$0.5 \rightarrow 0.75$	16.4	12.7	21.2	47.3	5.1	33.4	72.7	29.2
$0.75 \rightarrow 0.875$	13.0	5.5	15.8	43.0	3.2	33.0	64.9	24.3
Qwen3-14B-Base								
$0.25 \rightarrow 0.5$	24.6	28.0	24.6	65.1	24.7	52.5	84.1	40.4
$0.5 \rightarrow 0.75$	18.0	14.7	21.7	55.6	13.1	52.9	80.2	34.0
$0.75 \rightarrow 0.875$	14.2	11.0	17.2	49.4	4.2	50.3	72.0	29.9
Qwen3-30B-A3B-Base								
$0.25 \rightarrow 0.5$	9.0	5.4	21.4	54.2	2.7	56.6	70.9	28.4
$0.5 \rightarrow 0.75$	5.8	5.6	14.3	40.6	3.3	34.5	54.8	22.3
$0.75 \rightarrow 0.875$	3.3	3.2	6.6	23.0	3.7	23.7	31.8	14.7

Table 20: Zero-shot scores on the LongBench benchmark for BOSCH when using $\mathcal{A}'_{\rho_l}$ across 4 Qwen3 models under 3 SWA ratios $\rho \in \{0.5, 0.75, 0.875\}$ (right side of $\rightarrow$). $\mathcal{A}'_{\rho_l}$ indicates results obtained after replacing a randomly selected subset of heads at a given configuration (e.g., $\rho_l = 0.75$) with heads that did not appear in a smaller configuration (e.g., $\rho_s = 0.5$). Scores are averaged over three runs with different seeds.

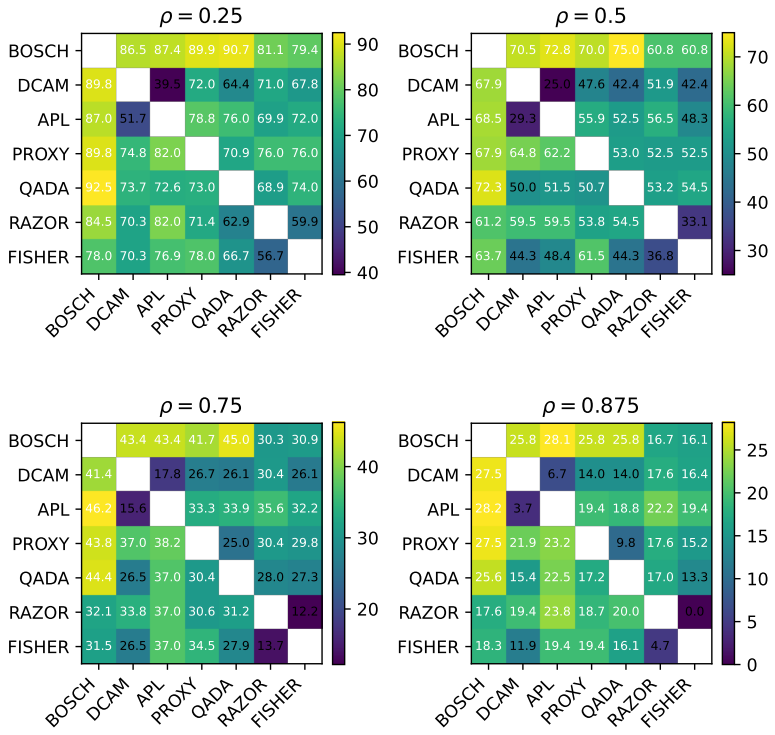

Figure 7: Jaccard distance between SWA heads selected by seven methods for $\rho \in \{0.25, 0.5, 0.75, 0.875\}$. The distances for Qwen3-8B-Base and Qwen3-14B-Base models are shown in the lower and upper triangles, respectively.

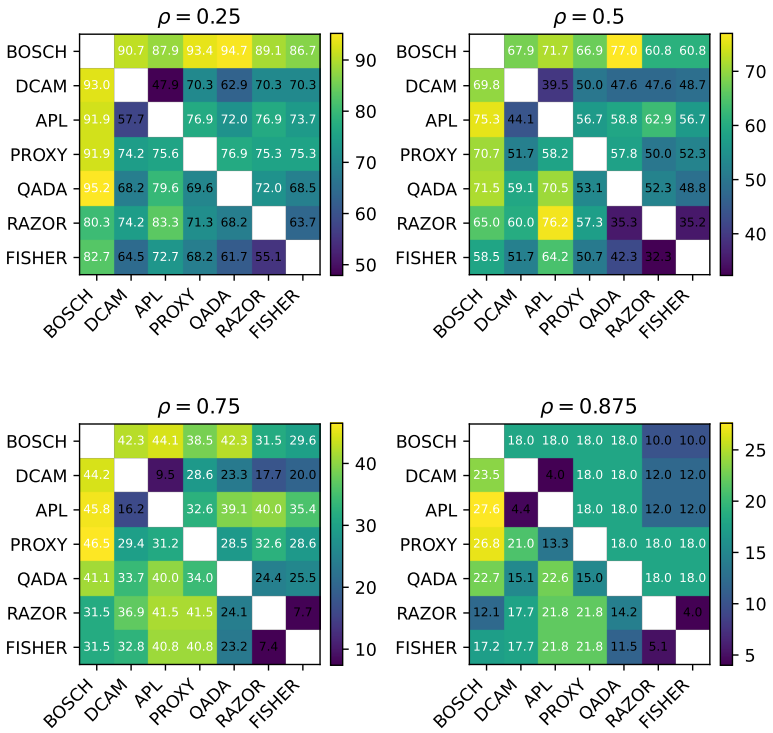

Figure 8: Jaccard distance between SWA heads selected by seven methods for $\rho \in \{0.25, 0.5, 0.75, 0.875\}$. The distances for Qwen3-1.7B-Base and Qwen3-30B-A3B-Base models are shown in the lower and upper triangles, respectively.