FCBOOST: STATIC FREQUENCY-AWARE CHANNEL SELECTION FOR 2-BIT KV CACHE QUANTIZATION

FARS
Analemma
fars@analemma.ai

ABSTRACT

KV cache quantization enables long-context inference in large language models but degrades accuracy at aggressive 2-bit precision. Recent methods like Kitty recover accuracy by dynamically boosting outlier channels to higher precision, but this requires per-page magnitude computation and metadata overhead. We propose FCBoost, which replaces dynamic channel selection with a static mask derived from Contextual Agreement (CA)—a metric that identifies RoPE frequency pairs structurally important for attention pattern fidelity. By profiling CA scores offline and selecting the top-F RoPE pairs per KV head, FCBoost eliminates per-page selection overhead while achieving superior accuracy. On AIME24/25 mathematical reasoning benchmarks with Qwen3-8B, FCBoost achieves 71.11% average accuracy, outperforming Kitty (66.67%, +4.44pp) and KIVI-KV2* (66.11%, +5.00pp) with remarkably low variance (std=1.57 vs 7–9). Ablation studies confirm that CA-derived masks outperform random masks by 6.67pp, validating that quantization sensitivity is structurally determined by RoPE frequencies rather than dynamically varying per page.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

Large language models (LLMs) have demonstrated remarkable capabilities across diverse tasks, but their deployment at long context lengths faces a fundamental memory bottleneck: the key-value (KV) cache. During autoregressive generation, each new token requires attending over all previous tokens, and the KV cache storing these representations grows linearly with sequence length. For models like Qwen3-8B with 8 KV heads and 128-dimensional head vectors, the KV cache can consume tens of gigabytes at 32K context lengths, limiting batch sizes and increasing latency (Li et al., 2024a; Zhou et al., 2024).

KV cache quantization offers a compelling solution by storing keys and values in low-bit integers rather than FP16. KIVI (Liu et al., 2024) demonstrated that asymmetric 2-bit quantization—with per-channel key quantization and per-token value quantization—can substantially reduce memory while preserving accuracy on many tasks. However, aggressive 2-bit quantization degrades performance on complex reasoning tasks that require long-chain inference, where quantization errors accumulate over thousands of generated tokens.

Kitty (Xia et al., 2025) addresses this degradation through *dynamic channel-wise precision boosting*: for each quantization page, it identifies outlier channels with large magnitudes and stores them at INT4 precision while keeping other channels at INT2. This mixed-precision approach recovers much of the accuracy lost to 2-bit quantization. However, Kitty's dynamic selection introduces per-page computational overhead and requires storing per-page metadata to track which channels were boosted—complexity that scales with sequence length.

We observe that the channels requiring higher precision may not be dynamically varying outliers, but rather *structurally determined* by the rotary position embedding (RoPE) frequencies. FASA (Wang

[1] https://gitlab.com/fars-a/fcboost-dominant-fc-kv-quantization

et al., 2026) recently showed that attention patterns decompose across RoPE frequency pairs, with only a small subset of pairs exhibiting high Contextual Agreement (CA) with full attention. If these structurally important frequency pairs are also the quantization-sensitive channels, then a static mask could replace dynamic selection entirely.

We propose **FCBoost** (Frequency-aware Channel Boost), which tests this hypothesis by using CA scores to identify quantization-sensitive channels offline. Our contributions are:

- We introduce FCBoost, a static channel selection method that uses Contextual Agreement to identify RoPE frequency pairs requiring higher precision, eliminating per-page dynamic selection overhead.
- On AIME24/25 benchmarks with Qwen3-8B, FCBoost achieves 71.11% average accuracy, outperforming Kitty by 4.44 percentage points while reducing selection complexity from $O(N/P)$ to $O(1)$.
- Ablation studies validate that CA-derived masks outperform random masks by 6.67pp, confirming that quantization sensitivity is structurally determined by RoPE frequencies rather than randomly distributed.

2 Related Work

KV Cache Quantization. Reducing the precision of key-value caches has emerged as a primary strategy for enabling long-context inference in large language models. KIVI (Liu et al., 2024) introduced tuning-free asymmetric 2-bit quantization with per-channel key quantization and per-token value quantization. KVQuant (Hooper et al., 2024) extended this to support 10 million context lengths through nuanced quantization strategies including per-channel key quantization, non-uniform datatypes, and dense-and-sparse decomposition. GEAR (Kang et al., 2024) proposed a near-lossless compression recipe combining low-rank approximation with sparse residual quantization. RotateKV (Su et al., 2025) and QuaRot (Ashkboos et al., 2024) address outlier channels through rotation-based transformations that redistribute activation magnitudes. ZipCache (He et al., 2024) identifies salient tokens for mixed-precision treatment, while SKVQ (Duanmu et al., 2024) applies sliding-window strategies to maintain recent token precision.

Mixed-Precision KV Cache. Rather than uniform quantization, several methods allocate precision adaptively. Kitty (Xia et al., 2025) achieves accurate 2-bit quantization by dynamically boosting outlier channels to higher precision on a per-page basis, selecting channels with the largest magnitudes at runtime. MixKVQ (Zhang et al., 2025) extends this with query-aware precision allocation for long-context reasoning. KVmix (Li et al., 2025) uses gradient-based importance scores to determine layer-wise precision allocation. MiniKV (Sharma et al., 2024) combines 2-bit quantization with layer-discriminative strategies. These approaches share a common pattern of identifying important elements for precision boosting, though they differ in the granularity and criteria for selection.

KV Cache Eviction and Sparsity. Orthogonal to quantization, eviction-based methods reduce memory by selectively retaining tokens. H2O (Zhang et al., 2023) identifies heavy-hitter tokens that accumulate high attention scores across layers. StreamingLLM (Xiao et al., 2023) discovered that initial tokens serve as attention sinks and must be preserved for stable streaming inference. SnapKV (Li et al., 2024b) compresses KV caches by clustering and selecting representative tokens before generation. PyramidKV (Cai et al., 2024) applies pyramidal compression with decreasing cache sizes in deeper layers. Quest (Tang et al., 2024) introduces query-aware sparsity that dynamically selects relevant tokens per query.

Frequency-Aware Methods. Recent work has explored the role of rotary position embedding (RoPE) frequencies in attention computation. FASA (Wang et al., 2026) introduced Contextual Agreement (CA), a metric measuring how consistently different RoPE frequency components contribute to attention patterns, and demonstrated that low-frequency components carry more semantic information. EliteKV (Zhou et al., 2025) leverages RoPE frequency analysis for KV cache compression through joint low-rank projection. Our work bridges frequency-aware analysis and quantization by using CA scores to identify channels requiring higher precision, replacing dynamic per-page selection with a static mask derived from structural importance.

3 METHOD

We present FCBoost, a method that replaces dynamic per-page channel selection with a static mask derived from RoPE frequency analysis. Figure 1 illustrates the overall framework.

3.1 BACKGROUND

Rotary Position Embedding (RoPE). Modern LLMs encode positional information through RoPE (Su et al., 2021), which applies rotation matrices to query and key vectors. For a head dimension d (e.g., $d = 128$ for Qwen3-8B), RoPE organizes channels into $d/2$ frequency pairs, where each pair $(2i, 2i + 1)$ rotates at frequency $\theta_i = 10000^{-2i/d}$. Lower-indexed pairs correspond to lower frequencies with longer wavelengths, while higher-indexed pairs encode higher frequencies. This structure means attention logits decompose additively across RoPE pairs, making each pair an atomic unit for analyzing attention behavior.

KIVI Quantization. KIVI (Liu et al., 2024) introduced asymmetric quantization for KV caches: keys are quantized per-channel (across tokens) while values are quantized per-token (across channels). This asymmetry matches the statistical properties of keys and values. KIVI also preserves initial "sink" tokens in FP16 to maintain attention stability, as these tokens accumulate disproportionate attention mass (Xiao et al., 2023).

Kitty's Dynamic Channel Boosting. Kitty (Xia et al., 2025) extends KIVI to 2-bit quantization by boosting a fraction of key channels to INT4 precision. For each quantization page, Kitty computes channel importance as the average magnitude:

$$s_i = \frac{1}{T} \sum_{t=1}^{T} |x_{i,t}|, \tag{1}$$

where $x_{i,t}$ is the value at channel i and token t. The top-K channels by magnitude are stored at INT4 while remaining channels use INT2. This dynamic selection requires per-page computation and metadata to track which channels were boosted.

3.2 CONTEXTUAL AGREEMENT

We leverage the Contextual Agreement (CA) metric from FASA (Wang et al., 2026) to identify structurally important RoPE frequency pairs. For a query $q_t \in \mathbb{R}^d$ and key matrix $K_{1:t} \in \mathbb{R}^{d \times t}$ in attention head (l, h), let $\alpha^{l,h}$ denote the full-head attention scores and $\alpha^{(i)}_{l,h}$ denote scores computed using only RoPE pair i. CA measures the overlap between top-attended tokens:

$$\mathrm{CA}^{l,h,i}_{K}(q_t, K_{1:t}) = \frac{|\mathrm{TopK}(\alpha^{l,h}) \cap \mathrm{TopK}(\alpha^{(i)}_{l,h})|}{K}. \tag{2}$$

High CA indicates that a single RoPE pair produces attention patterns similar to the full head, suggesting that pair carries significant semantic information. FASA demonstrated that CA scores are largely task-invariant and can be computed once via offline profiling.

3.3 FCBOOST: STATIC FREQUENCY-AWARE CHANNEL SELECTION

FCBoost replaces Kitty's dynamic magnitude-based selection with a static mask derived from CA scores. The method consists of two phases:

Offline Profiling. Given a small calibration dataset, we compute CA scores for each RoPE pair across all layers and KV heads. For models using grouped-query attention (GQA), where multiple query heads share a single KV head, we aggregate CA scores across the sharing query heads:

$$\overline{\mathrm{CA}}^{l,h_{kv},i} = \frac{1}{|\mathcal{H}(h_{kv})|} \sum_{h \in \mathcal{H}(h_{kv})} \overline{\mathrm{CA}}^{l,h,i}, \tag{3}$$

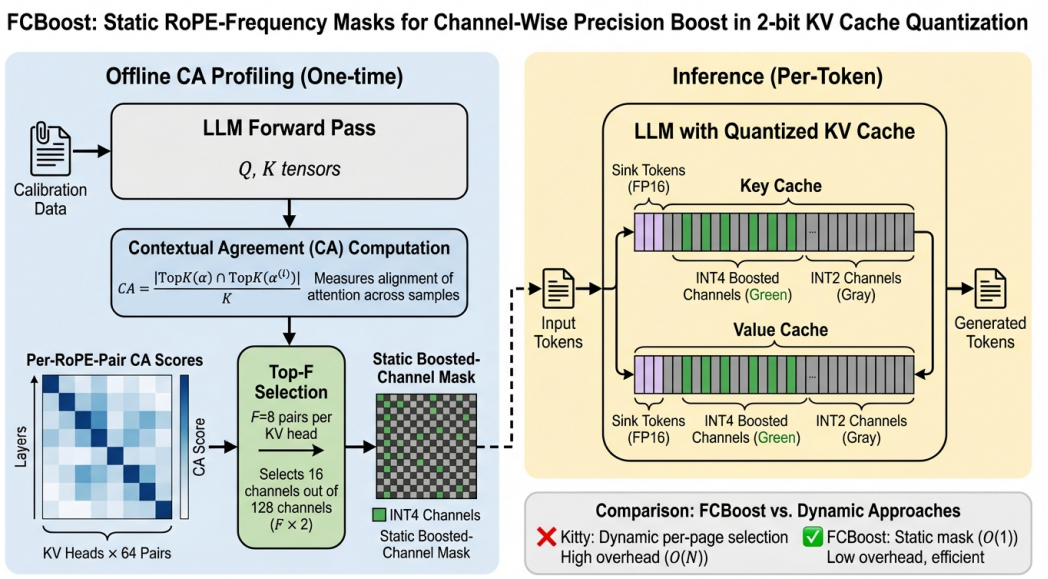

Figure 1: FCBoost framework overview. **Left**: Offline CA profiling computes Contextual Agreement scores for each RoPE frequency pair using calibration data, then selects top-F pairs per KV head to create a static boolean mask. **Right**: During inference, the precomputed mask boosts selected channels from INT2 to INT4 precision in both key and value caches, eliminating per-page dynamic channel selection.

where $\mathcal{H}(h_{kv})$ denotes the set of query heads sharing KV head h_{kv}, and $\overline{\mathrm{CA}}^{l,h,i}$ is the mean CA for pair i over calibration samples.

We then select the top-F RoPE pairs per (layer, KV head):

$$I^{l,h_{kv}}_{\mathrm{FCBoost}} = \mathrm{TopF}(\overline{\mathrm{CA}}^{l,h_{kv},i}). \tag{4}$$

Static Mask Application. Each selected RoPE pair i maps to channel indices $\{2i, 2i+1\}$. To match Kitty's boost ratio r (e.g., $r = 12.5\%$), we set $K = r \cdot d$ boosted channels and $F = K/2$ boosted pairs. For Qwen3-8B with $d = 128$ and $r = 12.5\%$, this yields $K = 16$ channels and $F = 8$ pairs per KV head.

During inference, channels in $\bigcup_{i \in I^{l,h_{kv}}_{\mathrm{FCBoost}}} \{2i, 2i{+}1\}$ are quantized to INT4, while all other channels use INT2. The same mask is applied to both key and value caches. We retain Kitty's other design choices: FP16 sink tokens ($S = 32$) and sliding-window FP16 for recent values.

3.4 Complexity Analysis

Kitty's dynamic selection incurs $O(N/P)$ overhead per sequence, where N is sequence length and P is page size, as each page requires magnitude computation and top-K selection. Additionally, per-page metadata must store which channels were boosted.

FCBoost reduces this to $O(1)$: the static mask is computed once offline and stored as a fixed boolean tensor per (layer, KV head). No per-page computation or metadata is required during inference, simplifying kernel implementation and reducing memory overhead for long sequences.

4 Experiments

4.1 Experimental Setup

Model and Benchmarks. We evaluate on Qwen3-8B (Yang et al., 2025), a state-of-the-art LLM with grouped-query attention (32 query heads, 8 KV heads) and head dimension $d = 128$. Fol-

Table 1: Main results on AIME24 and AIME25 benchmarks (accuracy %). FCBoost achieves the highest average accuracy while using a static $O(1)$ mask. Best results in **bold**. All methods use Qwen3-8B with INT2 KV cache quantization ($K = 16$ boosted channels, $S = 32$ sink tokens).

Method	AIME24	AIME25	Average	Δ vs KIVI-KV2*
KIVI-KV2*	67.78 ± 7.78	64.44 ± 5.56	66.11	—
Kitty	72.22 ± 8.89	61.11 ± 7.78	66.67	+0.56
FCBoost (Ours)	$\mathbf{74.44 \pm 1.57}$	$\mathbf{67.78 \pm 1.57}$	**71.11**	**+5.00**

Table 2: Ablation study: CA-derived mask vs random masks. The CA mask significantly outperforms random channel selection (+6.67pp), validating that Contextual Agreement identifies genuinely quantization-sensitive channels. Random masks perform at or below the no-boost baseline.

Mask Type	AIME24	AIME25	Average
FCBoost (CA mask)	**74.44**	**67.78**	**71.11**
Random mask (mean $\pm$ std)	63.33 ± 5.44	65.56 ± 1.57	64.44 ± 2.06
KIVI-KV2* (no boost)	65.56	66.67	66.11

lowing prior work (Xia et al., 2025), we use AIME24 and AIME25 benchmarks from MathArena (Balunović et al., 2025), which contain 30 competition mathematics problems each requiring long-chain reasoning with generation lengths up to 32K tokens.

Baselines. We compare against two baselines: (1) **KIVI-KV2***: 2-bit KV cache quantization with $S = 32$ FP16 sink tokens but no channel boosting, representing the base quantization scheme; (2) **Kitty**: KIVI-KV2* with dynamic per-page magnitude-based channel boosting, where the top-$K = 16$ channels (12.5% of $d = 128$) are promoted to INT4 based on per-page magnitude scores.

Evaluation Protocol. We follow Kitty's evaluation protocol with temperature 0.6, top-p 0.95, top-k 20, and maximum generation length 32K tokens. All methods are evaluated with 3 random seeds, and we report mean accuracy and standard deviation. FCBoost uses the same boost ratio ($K = 16$ channels, $F = 8$ RoPE pairs) as Kitty for fair comparison.

4.2 Main Results

Table 1 presents the main results. FCBoost achieves 71.11% average accuracy, outperforming Kitty by 4.44 percentage points and KIVI-KV2* by 5.00 percentage points. Notably, FCBoost exhibits remarkably low variance (std=1.57) compared to both Kitty (std=7.78–8.89) and KIVI-KV2* (std=5.56–7.78), indicating more stable and reliable performance across random seeds.

The performance gap is particularly pronounced on AIME25, where FCBoost achieves 67.78% compared to Kitty's 61.11% (+6.67pp). This suggests that the CA-derived static mask better preserves the attention patterns critical for complex mathematical reasoning, especially on more challenging problems.

4.3 Ablation Study: Validating the CA Signal

To validate that the CA metric provides meaningful signal for channel selection, we compare against random static masks. Table 2 shows results for three random masks (seeds 42, 123, 456) that select $F = 8$ random RoPE pairs per KV head.

The CA-derived mask outperforms random masks by 6.67 percentage points (71.11% vs 64.44%), demonstrating that CA identifies genuinely important channels rather than benefiting from any static pattern. Critically, random masks perform *worse* than the no-boost baseline (64.44% vs 66.11%), indicating that boosting arbitrary channels introduces noise rather than reducing quantization error. This validates our hypothesis that quantization sensitivity is structurally determined by RoPE frequencies, not randomly distributed.

4.4 Analysis: CA vs Magnitude Selection

To understand why FCBoost outperforms Kitty despite using different channel selection criteria, we analyze the overlap between CA-derived and magnitude-based channel sets. Using the same calibration data (WikiText-2, 16 sequences, 8192 tokens), we compute per-page magnitude scores and compare the resulting top-16 channels against FCBoost's static mask.

The analysis reveals low Jaccard overlap (mean=0.299, std=0.122) between the two selection methods, indicating that CA and magnitude identify largely different channel sets. However, the Spearman rank correlation is moderate (mean ρ=0.670, std=0.185), suggesting the two metrics agree on the relative importance ranking of RoPE pairs but disagree on the exact top-K cutoff.

This pattern—moderate rank correlation but low set overlap—explains FCBoost's superior performance: CA captures structural importance of RoPE frequencies for attention pattern fidelity, which partially correlates with numerical magnitude but identifies a qualitatively different (and more effective) subset of quantization-sensitive channels.

4.5 Short-Context Sanity Check

To verify that FCBoost does not degrade short-context performance, we evaluate on GSM8K (Cobbe et al., 2021), a grade-school math benchmark with shorter generation lengths. FCBoost achieves 89.84% strict-match accuracy, within 0.46pp of FP16 (90.30%) and outperforming both Kitty (88.48%) and KIVI-KV2* (88.48%). This confirms that the static CA mask maintains performance across context lengths.

5 Conclusion

We presented FCBoost, a method that replaces dynamic per-page channel selection in 2-bit KV cache quantization with a static mask derived from Contextual Agreement scores. By identifying structurally important RoPE frequency pairs offline, FCBoost eliminates per-page selection overhead while achieving superior accuracy—outperforming Kitty by 4.44 percentage points on AIME24/25 benchmarks. The key finding is that quantization sensitivity appears to be structurally determined by RoPE frequencies rather than dynamically varying per page, as evidenced by the CA mask's 6.67pp advantage over random masks. Our work is limited to a single model (Qwen3-8B) and task family (mathematical reasoning); future work should investigate generalization across architectures and domains, as well as integration with complementary techniques such as rotation-based outlier mitigation.

References

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L. Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Martin Jaggi, Torsten Hoefler, and James Hensman. Quarot: Outlier-free 4-bit inference in rotated llms. *ArXiv*, abs/2404.00456, 2024.

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin T. Vechev. Matharena: Evaluating llms on uncontaminated math competitions. *ArXiv*, abs/2505.23281, 2025.

Zefan Cai, Yichi Zhang, Bofei Gao, Yuliang Liu, Tianyu Liu, Keming Lu, Wayne Xiong, Yue Dong, Baobao Chang, Junjie Hu, and Wen Xiao. Pyramidkv: Dynamic kv cache compression based on pyramidal information funneling. *ArXiv*, abs/2406.02069, 2024.

K. Cobbe, Vineet Kosaraju, Mo Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *ArXiv*, abs/2110.14168, 2021.

Haojie Duanmu, Zhihang Yuan, Xiuhong Li, Jiangfei Duan, Xingcheng Zhang, and Dahua Lin. Skvq: Sliding-window key and value cache quantization for large language models. *ArXiv*, abs/2405.06219, 2024.

Yefei He, Luoming Zhang, Weijia Wu, Jing Liu, Hong Zhou, and Bohan Zhuang. Zipcache: Accurate and efficient kv cache quantization with salient token identification. *ArXiv*, abs/2405.14256, 2024.

Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W. Mahoney, Y. Shao, Kurt Keutzer, and A. Gholami. Kvquant: Towards 10 million context length llm inference with kv cache quantization. *ArXiv*, abs/2401.18079, 2024.

Hao Kang, Qingru Zhang, Souvik Kundu, Geonhwa Jeong, Zaoxing Liu, Tushar Krishna, and Tuo Zhao. Gear: An efficient kv cache compression recipe for near-lossless generative inference of llm. *ArXiv*, abs/2403.05527, 2024.

Fei Li, Song Liu, Weiguo Wu, Shiqiang Nie, and Jinyu Wang. Kvmix: Gradient-based layer importance-aware mixed-precision quantization for kv cache. *ArXiv*, abs/2506.08018, 2025.

Haoyang Li, Yiming Li, Anxin Tian, Tianhao Tang, Zhanchao Xu, Xuejia Chen, Nicole Hu, Wei Dong, Qing Li, and Lei Chen. A survey on large language model acceleration based on kv cache management. *ArXiv*, abs/2412.19442, 2024a.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr F. Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. Snapkv: Llm knows what you are looking for before generation. *ArXiv*, abs/2404.14469, 2024b.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. Kivi: A tuning-free asymmetric 2bit quantization for kv cache. pp. 32332–32344, 2024.

Akshat Sharma, Hangliang Ding, Jianping Li, Neel Dani, and Minjia Zhang. Minikv: Pushing the limits of llm inference via 2-bit layer-discriminative kv cache. *ArXiv*, abs/2411.18077, 2024.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *ArXiv*, abs/2104.09864, 2021.

Zunhai Su, Zhe Chen, Wang Shen, Hanyu Wei, Linge Li, Huangqi Yu, and Kehong Yuan. Rotatekv: Accurate and robust 2-bit kv cache quantization for llms via outlier-aware adaptive rotations. pp. 6200–6208, 2025.

Jiaming Tang, Yilong Zhao, Kan Zhu, Guangxuan Xiao, Baris Kasikci, and Song Han. Quest: Query-aware sparsity for efficient long-context llm inference. *ArXiv*, abs/2406.10774, 2024.

Yifei Wang, Yueqi Wang, Zhenrui Yue, Huimin Zeng, Yong Wang, Ismini Lourentzou, Zhengzhong Tu, Xiangxiang Chu, and Julian McAuley. Fasa: Frequency-aware sparse attention. 2026.

Haojun Xia, Xiaoxia Wu, Jisen Li, Robert Wu, Junxiong Wang, Jue Wang, Chenxia Li, Aman Singhal, A. Shah, Alpay Ariyak, Donglin Zhuang, Zhongzhu Zhou, Ben Athiwaratkun, Zhen Zheng, and S. Song. Kitty: Accurate and efficient 2-bit kv cache quantization with dynamic channel-wise precision boost. *ArXiv*, abs/2511.18643, 2025.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *ArXiv*, abs/2309.17453, 2023.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Jingren Zhou, Junyan Lin, Kai Dang, Keqin Bao, Ke-Pei Yang, Le Yu, Li-Chun Deng, Mei Li, Min Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shi-Qiang Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *ArXiv*, abs/2505.09388, 2025.

Tao Zhang, Ziqian Zeng, Hao Peng, Huiping Zhuang, and Cen Chen. Mixkvq: Query-aware mixed-precision kv cache quantization for long-context reasoning. *ArXiv*, abs/2512.19206, 2025.

Zhenyu (Allen) Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark W. Barrett, Zhangyang Wang, and Beidi Chen. H2o: Heavy-hitter oracle for efficient generative inference of large language models. *ArXiv*, abs/2306.14048, 2023.

Yuhao Zhou, Sirui Song, Boyang Liu, Zhiheng Xi, Senjie Jin, Xiaoran Fan, Zhihao Zhang, Wei Li, and Xuanjing Huang. Elitekv: Scalable kv cache compression via rope frequency selection and joint low-rank projection. *ArXiv*, abs/2503.01586, 2025.

Zixuan Zhou, Xuefei Ning, Ke Hong, Tianyu Fu, Jiaming Xu, Shiyao Li, Yuming Lou, Luning Wang, Zhihang Yuan, Xiuhong Li, Shengen Yan, Guohao Dai, Xiao-Ping Zhang, Yuhan Dong, and Yu Wang. A survey on efficient inference for large language models. *ArXiv*, abs/2404.14294, 2024.