

# DRAFT-BASED APPROXIMATE INFERENCE FOR LLMs

Kevin Galim<sup>1\*</sup> Ethan Ewer<sup>2\*</sup> Wonjun Kang<sup>1,3</sup> Minjae Lee<sup>1</sup>  
Hyung Il Koo<sup>1,4</sup> Kangwook Lee<sup>2,5</sup>

<sup>1</sup>FuriosaAI <sup>2</sup>UW-Madison <sup>3</sup>Seoul National University

<sup>4</sup>Ajou University <sup>5</sup>KRAFTON

{kevin.galim, kangwj1995, minjae.lee, hikoo}@furiosa.ai  
{eewer, kangwook.lee}@wisc.edu

Code: <https://github.com/furiosa-ai/draft-based-approx-llm>

## ABSTRACT

Optimizing inference for long-context large language models (LLMs) is increasingly important due to the quadratic compute and linear memory cost of Transformers. Existing approximate inference methods, including key-value (KV) cache dropping, sparse attention, and prompt compression, typically rely on coarse predictions of token or KV pair importance. We unify and extend recent work by introducing a framework for approximate LLM inference that leverages small draft models to more accurately predict token and KV pair importance. We provide novel theoretical and empirical analyses justifying lookahead-based importance estimation techniques. Within this framework, we present: (i) **SpecKV**, the first method to use lookahead with a small draft model to enable precise KV cache dropping; (ii) **SpecPC**, which leverages draft model attention activations to identify and discard less important prompt tokens; and (iii) **SpecKV-PC**, a cascaded compression strategy combining both techniques. Extensive experiments on long-context benchmarks demonstrate that our methods consistently achieve higher accuracy than existing baselines while retaining the same efficiency gains in memory usage, latency, and throughput.

## 1 INTRODUCTION

The demand for longer context lengths in large language models (LLMs) (Achiam et al., 2023; Google DeepMind, 2025) continues to grow (Liu et al., 2024b), driven by applications such as dialogue systems (Achiam et al., 2023; Google DeepMind, 2025), document summarization (Liu et al., 2020), and code completion (Du et al., 2024). Modern models like GPT-4 (Achiam et al., 2023) and Gemini 2.5 Pro (Google DeepMind, 2025) have pushed context windows to over a million tokens. However, scaling Transformers (Vaswani et al., 2017) to these lengths remains difficult due to significant computational and memory constraints. Attention computation scales quadratically with context length, increasing inference latency, while key-value (KV) cache memory grows linearly, straining GPU resources. For example, caching the KV states for 128K tokens in Llama-3.1-8B (Grattafiori et al., 2024) can consume over 16GB of memory, limiting the practical scalability of LLMs.

To address scalability challenges, recent work introduces approximate LLM inference techniques that reduce latency and memory usage at inference time. Techniques include sparse attention for prefilling (Jiang et al., 2024a) and decoding (Tang et al., 2024), which speed up inference by having each query attend to only a subset of keys. Sparse prefilling shortens time to the first token, while sparse decoding boosts generation throughput. KV cache dropping (Cai et al., 2024b; Li et al., 2024; Xiao et al., 2024; Zhang et al., 2023) reduces memory and increases throughput by shrinking the cache after prefilling or during decoding. Prompt compression (Choi et al., 2024; Jiang et al., 2023b; Liskavets et al., 2025) further improves efficiency by removing less important tokens before inputting the prompt, reducing both attention and MLP computation, as well as decreasing KV cache size.

---

\*Equal contribution.

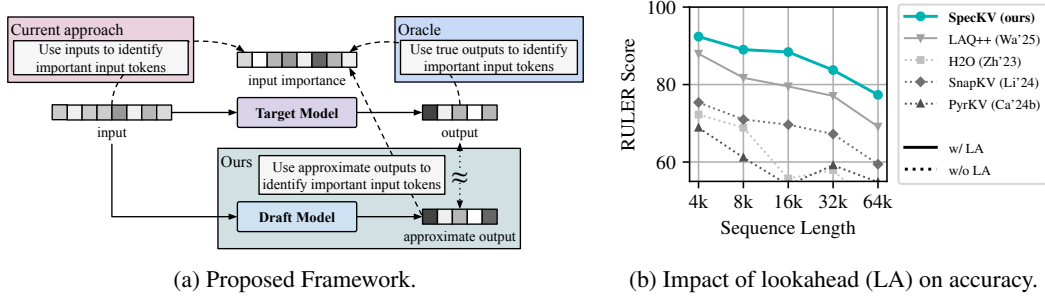


Figure 1: **(a)** Overview of our Draft-based Approximate Inference framework for input token importance estimation in comparison with current approaches and the oracle approach. Prior methods use input tokens to estimate input token importance. Our approach incorporates draft model predictions of future output tokens, yielding more accurate importance estimates. This better aligns with the hypothetical oracle setting, where the true output is known and influential tokens can be precisely identified. **(b)** On RULER with Llama-3-70B (Grattafiori et al., 2024), lookahead-based methods (LAQ++ (Wang et al., 2025), SpecKV) significantly outperform non-lookahead approaches (H2O (Zhang et al., 2023), SnapKV (Li et al., 2024), PyramidKV (Cai et al., 2024b)), with our proposed SpecKV achieving the best overall downstream score.

Orthogonally, speculative decoding (Cai et al., 2024a; Chen et al., 2023; Hu et al., 2025; Leviathan et al., 2023) accelerates LLM inference by using a small draft model to propose a sequence of multiple tokens, which the target model verifies in parallel. This improves throughput without altering the output distribution and is particularly effective for autoregressive models, where sequential generation is a bottleneck. However, unlike approximate inference, speculative decoding does not lower the total memory or computation requirements and struggles with increasing context length.

In contrast, approximate LLM inference improves efficiency by reducing the amount of computation the model performs. This is often done by estimating the importance of each token or KV pair for future generation and discarding less important ones from attention or feedforward computations. Existing methods (Cai et al., 2024b; Feng et al., 2025; Jiang et al., 2024a; Li et al., 2024) use attention activations from input tokens to predict which tokens or KV pairs future tokens will attend to, as future tokens are not yet available. However, input attention activations alone do not reliably identify the tokens or KV pairs most relevant for future token generation. Recent methods (Liu et al., 2025; Wang et al., 2025) address this by leveraging an approximate output to improve importance estimates. In this work, we unify and extend these techniques by introducing Draft-based Approximate Inference, a lookahead-based framework that uses an inexpensive draft model to approximate future outputs with minimal overhead (Fig. 1a). Our main contributions are as follows:

1. We present Draft-based Approximate Inference, a framework using draft model lookahead for enhanced approximate inference.
2. We present theoretical and empirical analyses justifying lookahead-based KV cache dropping and the use of draft model token importance to approximate target model token importance.
3. Within the Draft-based Approximate Inference framework, we develop two concrete algorithms targeting three LLM inference optimizations: Speculative KV Dropping (SpecKV) for KV cache dropping with sparse prefill, and Speculative Prompt Compression (SpecPC) for prompt compression. Notably, SpecKV is the first to use draft model lookahead for KV cache optimization. Additionally, we introduce SpecKV-PC, a cascaded pipeline integrating both algorithms to achieve superior accuracy, latency, and memory efficiency.
4. We perform comprehensive experiments on long-context benchmarks, demonstrating that our methods attain state-of-the-art accuracy under fixed KV cache or prompt size constraints. Our results consistently outperform prior baselines, underscoring the potential of draft models for fast and accurate approximate inference in large language models.

Table 1: Summary of prior work. Complexity reported without auxiliary/draft model.  $n_{in}$  and  $n_{out}$  denote the number of input and output tokens, respectively.  $C_{max}$  represents the maximum KV cache or prompt capacity.  $s_{prefill}$  and  $s_{decode}$  indicate the number of keys each query attends to during the prefill and decoding phases, respectively.  $L$  is the number of attention layers. Since all time complexities are linear with respect to  $L$ , we only include  $L$  for space complexity. Highlighted cells indicate improved complexity.

Type	Method	Sparse Attn	KV Dropping	Prefill Time	Decoding Time	Prefill Space	Decoding Space
Dense	Dense	✗	✗	$O(n_{in}^2)$	$O(n_{out}(n_{in} + n_{out}))$	$O(Ln_{in})$	$O(L(n_{in} + n_{out}))$
Sparse attention	MInference, FlexPrefill	Prefill	✗	$O(n_{in}s_{prefill})$	$O(n_{out}(n_{in} + n_{out}))$	$O(Ln_{in})$	$O(L(n_{in} + n_{out}))$
	Quest, RetrievalAttention	Decode	✗	$O(n_{in}^2)$	$O(n_{out}s_{decode})$	$O(Ln_{in})$	$O(L(n_{in} + n_{out}))$
KV dropping	StreamingLLM	Prefill	Decode	$O(n_{in}C_{max})$	$O(n_{out}C_{max})$	$O(\max(n_{in}, LC_{max}))$	$O(LC_{max})$
	H2O	✗	Decode	$O(n_{in}^2)$	$O(n_{out}C_{max})$	$O(\max(n_{in}, LC_{max}))$	$O(LC_{max})$
	SnapKV, PyramidKV, AdaKV	✗	After prefill	$O(n_{in}^2)$	$O(n_{out}(C_{max} + n_{out}))$	$O(\max(n_{in}, LC_{max}))$	$O(L(C_{max} + n_{out}))$
	LAQ++	✗	After prefill	$O(n_{in}^2)$	$O(n_{out}(C_{max} + n_{out}))$	$O(Ln_{in})$	$O(L(C_{max} + n_{out}))$
	<b>SpecKV (Ours)</b>	Prefill	After prefill	$O(n_{in}s_{prefill})$	$O(n_{out}(C_{max} + n_{out}))$	$O(\max(n_{in}, LC_{max}))$	$O(L(C_{max} + n_{out}))$
Prompt compression	LLMLingua-2, CPC, R2C, SpecPrefill, <b>SpecPC (Ours)</b>	–	–	$O(C_{max}^2)$	$O(n_{out}(C_{max} + n_{out}))$	$O(LC_{max})$	$O(L(C_{max} + n_{out}))$

## 2 RELATED WORK

**Sparse Attention** One way to improve inference efficiency is through sparse attention with static patterns. For example, sliding window attention (Beltagy et al., 2020), used in models like Mistral 7B (Jiang et al., 2023a), Gemma 3 (Gemma Team, 2025), GPT-3 (Brown et al., 2020), and gpt-oss (Agarwal et al., 2025), restricts each query to attend only a fixed-size window of recent keys, reducing computation and KV cache size during decoding. StreamingLLM (Xiao et al., 2024) improves on sliding window by using initial tokens, called attention sinks, along with the sliding window. MInference (Jiang et al., 2024a), adopted by Qwen2.5-1M (Yang et al., 2025b), further boosts prefill efficiency by searching offline for adaptive sparse attention patterns (A-shape, Vertical-Slash, and Block-Sparse) assigned per head. FlexPrefill (Lai et al., 2025) extends this idea by determining sparsity rates for each input prompt. In contrast, Quest (Tang et al., 2024) and RetrievalAttention (Liu et al., 2024a) target the decoding stage by only retrieving the most important KV pairs from the cache, reducing both memory bandwidth and computational demands during generation.

**KV Cache Dropping** KV dropping reduces computation and memory during decoding. Sliding window attention (Beltagy et al., 2020) and StreamingLLM (Xiao et al., 2024) are examples of KV dropping methods (as well as sparse attention) as they permanently evict KV pairs from cache. H2O (Zhang et al., 2023) improves on this by dynamically selecting attention sinks, termed heavy-hitters, using attention scores at each decoding step, while also maintaining a sliding window. SnapKV (Li et al., 2024) compresses the KV cache at the end of the prefill stage by dropping unimportant KV pairs. Subsequent work extends this idea by allocating KV cache budgets non-uniformly across layers (PyramidKV (Cai et al., 2024b)) and attention heads (AdaKV (Feng et al., 2025), HeadKV (Fu et al., 2025)). However, these approaches drop tokens based only on current information, making them less robust to changes in token importance over time (Nawrot et al., 2025). Recently, Wang et al. (2025) proposed Lookahead Q-Cache (LAQ++), which addresses this by generating draft queries with a sparse approximation of the target model, using them to compute more accurate importance scores, though this comes at the cost of no reduction in peak memory usage.

**Prompt Compression** Prompt compression removes tokens before reaching the model, reducing compute and memory usage during both prefill and decoding, unlike KV dropping, which speeds up only decoding. It also surpasses sparse attention by saving both attention and MLP computation. Prompt compression works seamlessly with all types of inference setups, such as APIs or inference engines like vLLM (Kwon et al., 2023), since it does not require any modifications to the model. However, KV dropping can achieve higher compression because it selects tokens per head, while prompt compression drops the same tokens across all layers and heads.

In a question-answer setup, prompt compression may be question-agnostic (compressing context without considering the question) or question-aware (factoring in the question). Selective context (Li et al., 2023) and LLMLingua (Jiang et al., 2023b) are training-free, question-agnostic approaches using a small LLM to keep only key tokens. LongLLMLingua (Jiang et al., 2024b) adapts this

for longer contexts in a question-aware manner. LLMingua-2 (Pan et al., 2024) trains a small model (Conneau et al., 2020) to score token importance without using the question. CPC (Liskavets et al., 2025) uses a trained encoder to compute sentence importance via cosine similarity with the question, while R2C (Choi et al., 2024) splits the prompt into chunks, processes each with the question using a fine-tuned encoder-decoder Transformer (FiD (Izacard & Grave, 2021)), and ranks them via cross-attention. Similar to our proposed SpecPC, SpecPrefill (Liu et al., 2025) leverages attention scores from a smaller draft model to identify important tokens, using the draft model for lookahead to improve token importance estimates.

**Speculative Decoding** Speculative decoding (Cai et al., 2024a; Chen et al., 2023; Hu et al., 2025; Leviathan et al., 2023) accelerates LLM inference by using a small draft model to propose multiple tokens that the target model verifies in parallel. This increases decoding throughput without changing the output distribution, addressing the bottleneck of autoregressive generation. Previous work further accelerates speculative decoding by enabling approximate inference in the draft model, using techniques such as sparse attention (Sadhukhan et al., 2025), KV cache dropping (Sun et al., 2024), or KV cache quantization (Tiwari et al., 2025), all while preserving exact inference. In contrast, our approach leverages draft models to enable fast, approximate inference directly in the target model.

Table 1 summarizes prior work, highlighting their prefill, decoding time, and memory complexities.

### 3 PROPOSED FRAMEWORK: DRAFT-BASED APPROXIMATE INFERENCE

Previous LLM approximation methods (Cai et al., 2024b; Feng et al., 2025; Jiang et al., 2024a; Li et al., 2024) estimate the importance of current tokens on future generation by analyzing current attention patterns. While this can be effective, it provides only a rough estimate of each token’s importance. In contrast, if future tokens were available, we could make substantially better importance estimates by directly identifying which input tokens contribute to generating those output tokens. However, these future tokens are inaccessible before generation.

Recent work has explored using approximate future information to improve token importance estimation. LAQ++ (Wang et al., 2025) extends SnapKV (Li et al., 2024) by generating draft queries and then using them to compute more accurate importance scores. SpecPrefill (Liu et al., 2025) generates lookahead tokens with a small draft model and relies on the draft model’s attention to those tokens to estimate input token importance. We bring these ideas together under a unified framework, Draft-based Approximate Inference that leverages approximate future information to improve token importance estimation. We further extend this framework with two new algorithms: SpecKV for KV cache dropping and sparse prefilling (Section 4.1) and SpecPC for prompt compression (Section 4.2). Finally, we introduce cascaded compression with SpecKV-PC, combining both approaches into a single pipeline for superior accuracy and efficiency (Section 4.3).

While our methods use draft models, a technique also common in speculative decoding, our objective is fundamentally different. Speculative decoding improves hardware utilization by having a draft model propose tokens that the target model verifies, accelerating generation without changing the output distribution. However, it does not reduce total computation or memory usage. In contrast, our framework reduces computation and memory costs by approximating the target model.

#### 3.1 JUSTIFICATION FOR LOOKAHEAD-BASED KV CACHE DROPPING

KV cache dropping requires estimating the importance of each input KV pair. We define importance as the average attention activation from output queries to each input key. Specifically, the vector of importance scores and its approximation are given by

$$s^T = \frac{1}{n_{\text{out}}} \sum_{i=1}^{n_{\text{out}}} \text{Softmax} \left( \frac{x_i^{(o)T} W_q W_k^T X^T}{\sqrt{d}} \right) \quad \text{and} \quad \hat{s}^T = \frac{1}{n_{\text{out}}} \sum_{i=1}^{n_{\text{out}}} \text{Softmax} \left( \frac{\hat{x}_i^{(o)T} W_q W_k^T X^T}{\sqrt{d}} \right), \quad (1)$$

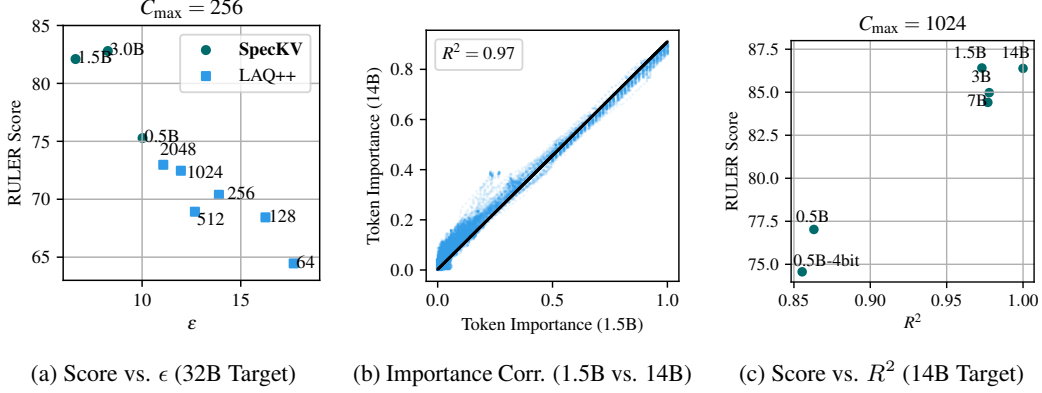


Figure 2: Experimental validation on RULER-32K tasks (5 samples each) using Qwen2.5 models. **(a)** Lower error  $\epsilon$  (Eq. (2)) yields higher downstream scores. Increasing the draft model size (SpecKV) or initial cache size (LAQ++)<sup>1</sup> reduces  $\epsilon$ , with SpecKV outperforming LAQ++. **(b)** Importance scores (as used in SpecPC) of the draft and target models are highly correlated. **(c)** For SpecPC, a larger draft model improves both the token importance correlation ( $R^2$ ) and the final task performance.

where  $X = [x_1, \dots, x_{n_{in}}]^T \in \mathbb{R}^{n_{in} \times d}$  is the matrix of input embeddings,  $x_i^{(o)} \in \mathbb{R}^d$  is the input embedding from the  $i$ th output token, and  $\hat{x}_i^{(o)} \in \mathbb{R}^d$  is its approximation.  $s_i$  and  $\hat{s}_i$  denote the importance of the  $i$ th KV pair.

To understand when lookahead-based KV cache dropping algorithms provide reliable importance estimates, we analyze how error in approximate input embeddings impacts error in importance score estimates for a single attention head.

**Theorem 1.** *If  $\|x_i^{(o)} - \hat{x}_i^{(o)}\|_2 \leq \epsilon$  for all  $i$  and  $\|x_j\|_2 \leq \sqrt{d}$  for all  $j$ , then  $\|s - \hat{s}\|_2 \leq \epsilon \|W_q W_k^T\|_2$ .*

This result shows that for a single attention layer, the worst-case error in the approximate importance scores is proportional to the worst-case error in the approximate input embeddings, implying that lookahead-based KV cache dropping algorithms provide reliable importance estimates as long as the draft model remains reasonably accurate (see Section B.1 for proof).

To evaluate the quality of different draft outputs, we compute  $\epsilon$  for two algorithms (LAQ++ and SpecKV). While Theorem 1 assumes a strict token-wise alignment (where each draft token  $\hat{x}_i^{(o)}$  corresponds to a target token  $x_i^{(o)}$ ), in practice, the draft and target models often generate sequences of different lengths ( $n_{out} \neq \hat{n}_{out}$ ). To address this mismatch, we relax the strict point-wise condition and empirically evaluate draft quality using the global distance between the centroids of the output sequences. We compute  $\epsilon$  as:

$$\epsilon = \left\| \frac{1}{n_{out}} \sum_{i=1}^{n_{out}} x_i^{(o)} - \frac{1}{\hat{n}_{out}} \sum_{i=1}^{\hat{n}_{out}} \hat{x}_i^{(o)} \right\|_2. \quad (2)$$

This metric serves as a practical proxy for the distributional mismatch between the target and draft outputs. We report  $\epsilon$  averaged across all attention layers. As shown in Fig. 2a, increasing draft quality (decreasing  $\epsilon$ ) leads to higher downstream accuracy. Furthermore, Section E.4 demonstrates how SpecKV’s lookahead mechanism improves importance score correlation and accuracy, particularly for long outputs.

### 3.2 JUSTIFICATION FOR DRAFT-BASED PROMPT COMPRESSION

Most KV cache dropping algorithms estimate token importance using attention scores from the target model itself. While effective, this approach is too computationally expensive for tasks like prompt compression, where we need to estimate token importance without forwarding the entire

<sup>1</sup>For LAQ++, initial cache size refers to the value of  $C_{max}$  used during the sparse draft generation.

input through the target model. A more efficient alternative is to use a smaller draft model for this estimation. To justify this approach, we study how the similarity between the draft and target models' outputs correlates with the similarity of their attention activations in a single attention layer.

The target model attention layer uses weights  $W_q$ ,  $W_k$ , and  $W_v$ , and the draft model attention layer uses  $\hat{W}_q$ ,  $\hat{W}_k$ , and  $\hat{W}_v$ . Let the input prompt be  $X = [x_1, \dots, x_n]^T \in \mathbb{R}^{n \times d}$ . The outputs of the target attention layer and its approximation are

$$Y = \text{Softmax}\left(\frac{XW_qW_k^TX^T}{\sqrt{d}}\right)XW_v = AXW_v, \quad \hat{Y} = \text{Softmax}\left(\frac{X\hat{W}_q\hat{W}_k^TX^T}{\sqrt{d}}\right)X\hat{W}_v = \hat{A}X\hat{W}_v, \quad (3)$$

where  $X = [x_1, \dots, x_n]^T \in \mathbb{R}^{n \times d}$ ,  $A = [a_1, \dots, a_n]^T$  is the attention matrix, and  $\hat{A} = [\hat{a}_1, \dots, \hat{a}_n]^T$  is the approximate attention matrix.

If the scaled inputs satisfy the Restricted Isometry Property (RIP)<sup>2</sup> (Candes & Tao, 2005), a condition widely studied in compressed sensing to ensure the stable recovery of sparse signals, we can establish the following bound:

**Theorem 2.** *If there exists a constant  $c$  such that  $cX^T$  satisfies the Restricted Isometry Property with parameters  $(2k, \delta)$ , where  $\delta$  is the restricted isometry constant and  $k$  is the approximate sparsity of  $a_i$  and  $\hat{a}_i$ , and the output error satisfies  $\|y_i - \hat{y}_i\|_2 \leq \epsilon \|X\|_{\infty, 2}$ , then the attention error satisfies*

$$\|a_i - \hat{a}_i\|_2 \leq \frac{2c\epsilon \|X\|_{\infty, 2}}{\sigma_{\min}(W_v)(1-\delta)}.$$

This result offers a surprising and elegant connection: it reveals that mathematical tools developed for compressed sensing can also bound the error in attention approximations. Specifically, it shows that the worst-case error in the approximate attention activations is proportional to the worst-case error in the approximate outputs, with the constant depending on the conditioning of the weight matrices and the maximum input embedding norm. This implies that if the draft model provides a reasonable approximation of the output, it also gives a reasonable approximation of the attention activations (see Section B.2 for proof). Furthermore, even if the scaled inputs do not satisfy the RIP, we can still bound the attention approximation error by applying Theorem 3 (see Section B.3 for proof).

In addition to the theoretical analysis, we examine the correlation between the importance scores (as computed by SpecPC) for Qwen2.5-Instruct (1.5B as draft, 14B as target). In Fig. 2b, we plot the draft importance scores against the corresponding target importance scores. The results reveal a strong correlation, supporting the use of draft attention activations to approximate target token importance. Furthermore, this correlation strengthens as the draft model size increases (Fig. 2c).

## 4 DRAFT-BASED APPROXIMATE INFERENCE METHODS

### 4.1 SPECKV: ROBUST IMPORTANCE ESTIMATION FOR KV CACHE DROPPING

Existing sparse attention and KV cache dropping methods, such as SnapKV, estimate token importance by analyzing recent attention activations. This approach can be inaccurate when the set of important KV pairs shifts during generation, as past patterns do not always predict future ones. We argue that a more robust estimate can be derived from the attention activations of draft queries for future tokens.

LAQ++ attempts this by generating draft queries from the target model using an initially compressed cache. These queries are then used to compute more accurate importance scores for a second, more informed compression pass. However, this two-pass method provides no reduction in peak memory because it must store the entire original KV cache of the target model to avoid recomputing the expensive prefill stage.

To overcome this limitation, we propose SpecKV. Our method employs a lightweight draft model to generate the draft output, which substantially reduces the cost of the lookahead step. This enables an accurate compression of the KV cache without sacrificing peak memory reduction.

<sup>2</sup>The input embedding matrix may satisfy the RIP if its entries are approximately uniformly or normally distributed. RIP can also hold with positional embeddings constructed from a Fourier basis.

<sup>3</sup> $\|X\|_{\infty, 2}$  denotes the maximum  $\ell_2$  norm of  $X$ 's rows;  $\sigma_{\min}(W_v)$  is the smallest singular value of  $W_v$ .

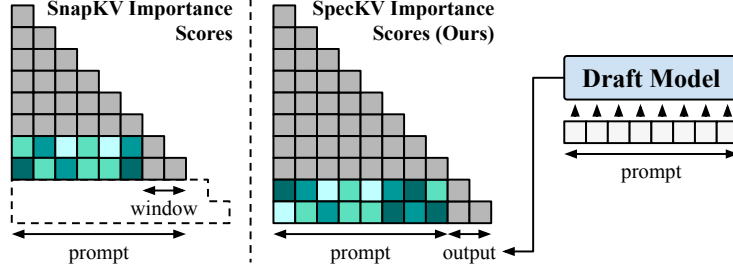


Figure 3: Overview of SpecKV: Instead of using only the last prompt tokens like SnapKV, SpecKV employs a lightweight draft model to generate lookahead tokens, providing richer context for more accurate KV importance estimation. Tokens in window are always retained.

SpecKV (Algorithm 1) begins by generating a draft output of length  $n_{\text{lookahead}}$  using a small draft model, which acts as a proxy for the target model’s future outputs. During prefilling, both the input tokens and the draft tokens are passed through the target model. For each attention head, we compute token importance scores by measuring the cross-attention between the queries from the last  $n_{\text{window}}$  input tokens and the draft output tokens to the remaining input keys (Fig. 3). We apply local pooling with kernel size  $k$  to the attention scores to maintain continuity. These scores guide two optimizations: sparse prefilling and KV cache dropping. For sparse prefilling, we use a variation of the Vertical-Slash kernel pattern introduced in (Jiang et al., 2024a). For KV cache dropping, we retain the top  $C_{\text{max}} - n_{\text{window}}$  KV pairs with the highest importance scores, along with the final  $n_{\text{window}}$  KV pairs from the most recent tokens.

#### 4.2 SPECPC: LEVERAGING DRAFT MODELS FOR EFFICIENT PROMPT COMPRESSION

SpecKV leverages the draft model outputs to enable more effective KV cache dropping. However, greater efficiency gains are possible by leveraging more information from the draft model. As demonstrated in Section 3.2, draft model attention scores serve as reliable estimates of target token importance. Building on this insight, we introduce SpecPC, an extension of SpecPrefill (Liu et al., 2025), which compresses the prompt to reduce latency and memory usage during both prefilling and decoding, surpassing the efficiency benefits of traditional KV cache dropping.

SpecPC (Algorithm 2) feeds an input prompt (length  $n_{\text{in}}$ ) to the draft model and directly extracts its attention activations  $A \in \mathbb{R}^{n_{\text{layer}} \times n_{\text{head}} \times (n_{\text{in}} + n_{\text{lookahead}} - 1) \times n_{\text{in}}}$ , where  $n_{\text{layer}}$  and  $n_{\text{head}}$  denote the number of layers and heads. These activations indicate token importance and are used to drop less relevant tokens from the prompt.

SpecPrefill (Liu et al., 2025) uses a window size of  $n_{\text{window}} = 1$ , meaning it relies on attention scores from queries associated with the last input token and  $n_{\text{lookahead}}$  draft output tokens to compute token importance. However, our experiments show that the optimal choice of  $n_{\text{window}}$  (the number of input queries used for importance estimation) is task-dependent, with some tasks benefiting from additional input queries. To address this, we adopt a large window with non-uniform token weights, placing greater emphasis on tokens near the prompt’s end to achieve robust, task-wide performance. Window tokens are reweighted so that the  $j$ th token from the end receives weight  $\frac{n_{\text{window}} - (j-1)}{n_{\text{window}}}$ . Max aggregation is performed across layers, heads, and queries to produce a single importance score per token (excluding the always-kept last  $n_{\text{window}}$  tokens). Additionally, as PyramidKV (Cai et al., 2024b) showed that attention is more focused in deeper layers, we exclude the first  $l_{\text{skip}}$  layers during aggregation.

We apply average, then max pooling, so selected tokens also include nearby context, avoiding static chunking that could split related tokens (e.g., key-value pairs). This maintains the local context LLMs require. Unlike other methods that select entire sentences, we avoid sentence-level pre-processing to support non-text inputs, such as images. We then select the top- $C_{\text{max}}$  tokens with the highest scores, always including window tokens, to form the compressed prompt. We pass the compressed prompt to the target model with reassigned contiguous position IDs, enabling SpecPC to surpass the model’s maximum context length.



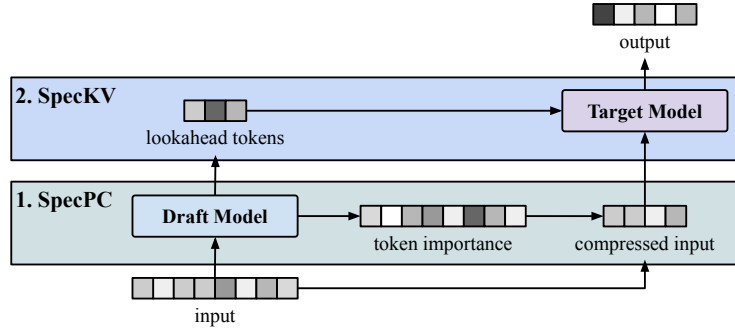


Figure 4: Overview of cascaded compression with SpecKV-PC: First, the draft model produces token importance scores and lookahead tokens. Next, SpecPC uses these scores to compress the initial input prompt. Finally, the target model is prefilled using both the compressed prompt and the lookahead tokens, while SpecKV compresses its KV cache

#### 4.3 SPECKV-PC: CASCADED COMPRESSION WITH SPECKV AND SPECPC

SpecKV-PC integrates SpecKV and SpecPC into a highly efficient, two-stage compression pipeline. The core strategy leverages a cascaded approach, which first compresses the prompt with SpecPC and then further compresses the KV cache with SpecKV (Fig. 4). Because of fewer target model activations, SpecKV-PC achieves substantially lower latency and a smaller memory footprint than SpecKV alone, as the computationally intensive target model processes only a fraction of the original prompt.

## 5 EXPERIMENTS

### 5.1 SETUP

We benchmark SpecKV and SpecPC against baselines on RULER (Hsieh et al., 2024) and LongBench (Bai et al., 2024) using two model pairs: Qwen2.5-Instruct (Yang et al., 2024) (14B target with 1.5B/0.5B drafts for KV dropping/prompt compression) and Llama-3-Instruct (Grattafiori et al., 2024) (3.1-70B 4-bit target with 3.2-3B/1B drafts, respectively).

RULER is a synthetic benchmark with 13 tasks of varying complexity, including tasks such as key-value retrieval (NIAH), multi-hop tracing, and aggregation. It can be generated at any sequence length to assess a model’s effective context window. We evaluate at 4K, 8K, 16K, 32K, and 64K (Qwen is excluded at 64K due to its 32K sequence limit). LongBench contains 12 English, five Chinese, two code, and three synthetic tasks across five categories. We exclude the Chinese tasks (unsupported by Llama) and synthetic tasks (already covered by RULER). We select 50 examples from each task, yielding 700 examples from LongBench and 650 examples at each context length for RULER.

For SpecKV, we compare against KV dropping methods H2O (Zhang et al., 2023), SnapKV (Li et al., 2024), PyramidKV (Cai et al., 2024b), and LAQ++ (Wang et al., 2025), using  $C_{\max} = 256$ . Additionally, we assess SpecKV-PC-2048 that first compresses to 2048 tokens via SpecPC before applying SpecKV ( $C_{\max} = 256$ ). For SpecPC, we benchmark against LLMingua-v2 (Pan et al., 2024), CPC (Liskavets et al., 2025), R2C (Choi et al., 2024), and SpecPrefill (Liu et al., 2025) with  $C_{\max} = 1024$ . Based on our ablation studies (Fig. 19), we set  $n_{\text{lookahead}}$  to the maximum token limit for SpecKV and to one for SpecPC. For SpecPrefill and LAQ++, we use  $n_{\text{lookahead}} = 8$  as in the official paper. Sections C to E contain datasets, details, and additional results, including cross-family evaluations, different  $C_{\max}$  settings, and multimodal experiments on MileBench (Dingjie et al., 2024) with Qwen2.5-VL (Bai et al., 2025).



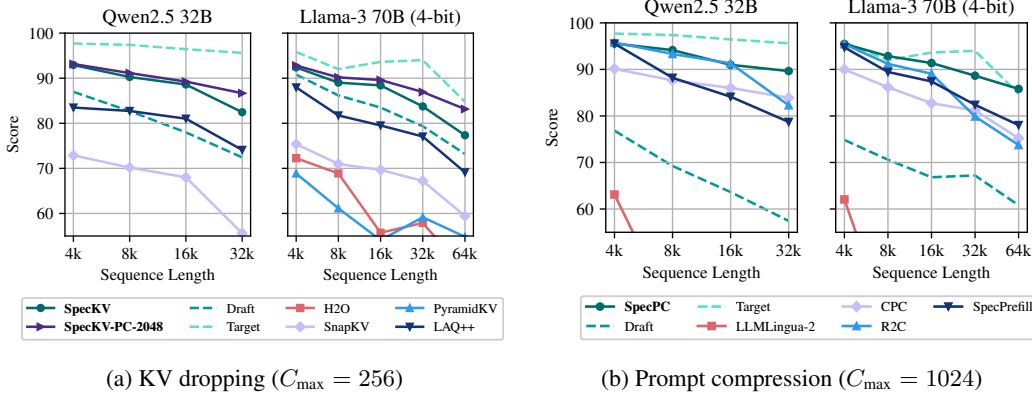


Figure 5: Performance of SpecKV and SpecPC. Both methods consistently outperform all baselines across sequence lengths, maintaining strong results at longer contexts. SpecKV-PC further improves upon SpecKV to achieve state-of-the-art results for KV dropping. Note that H2O and PyramidKV are not plotted for Qwen2.5 32B as their performance falls outside the visible range.

## 5.2 RESULTS

Fig. 5 (RULER) and Table 2 (LongBench) compare our methods with baselines. All our methods consistently outperform other baselines, demonstrating superior KV cache and prompt compression. Their performance far exceeds the draft model, highlighting robustness even with weaker drafts. Performance improves further with better drafts (Fig. 18). On RULER and LongBench, SpecKV and SpecPC consistently surpass existing baselines across multiple model families and context lengths. Interestingly, SpecKV-PC outperforms SpecKV alone, suggesting that the initial SpecPC-based prompt compression acts as an effective pre-filter by removing easy-to-identify, unimportant tokens. For larger  $C_{\max}$ , our methods remain superior (Figs. 10 and 11; Tables 9 to 11). Finally, we provide additional results regarding AdaKV (Section E.2), multimodal tasks (Section E.3), cross-family settings (Section E.5), and SpecKV-PC prompt compression ratios (Section E.6).

Table 2: LongBench performance with Qwen2.5 and Llama-3.

Group	Method	Qwen2.5 32B						Llama-3 70B (4-bit)					
		SingleQA	MultiQA	Summ.	Few-shot	Code	All	SingleQA	MultiQA	Summ.	Few-shot	Code	All
Dense	Target	56.01	43.99	25.90	64.06	44.74	47.78	55.02	47.06	28.61	70.47	48.19	49.99
KV	H2O	46.63	30.81	19.88	56.03	39.27	39.32	54.07	41.30	22.55	49.10	54.14	43.52
	SnapKV	52.54	40.21	19.89	61.18	40.12	42.98	<u>55.88</u>	45.30	22.49	62.15	55.49	47.75
	PyramidKV	50.92	37.26	18.90	63.24	40.20	43.19	55.41	45.59	22.50	59.06	49.90	46.25
	LAQ++	<b>55.15</b>	<u>44.14</u>	22.24	63.25	41.19	45.79	54.90	46.48	22.83	<u>64.31</u>	55.10	48.43
	SpecKV	<u>53.48</u>	43.77	<u>24.02</u>	<b>63.79</b>	<u>44.80</u>	<u>46.06</u>	51.80	<b>47.23</b>	<u>25.53</u>	64.02	<b>58.75</b>	<u>48.80</u>
	SpecKV-PC-2048	52.60	<b>44.52</b>	<b>24.11</b>	<u>63.38</u>	<b>48.45</b>	<b>46.48</b>	<b>61.42</b>	<u>47.15</u>	<b>26.51</b>	<b>66.94</b>	58.19	<b>51.60</b>
PC	LLMLingua-2	33.83	26.39	22.85	32.46	<u>43.01</u>	30.90	37.95	28.20	23.35	42.37	37.63	33.63
	CPC	45.60	40.62	23.09	60.08	32.31	40.91	45.14	39.41	24.86	61.40	37.58	41.97
	R2C	50.49	40.37	<u>23.26</u>	53.45	34.11	39.88	48.93	42.01	25.38	58.91	40.19	43.29
	SpecPrefill	45.94	39.32	23.16	62.04	<b>43.17</b>	<u>42.70</u>	<u>54.62</u>	<b>46.43</b>	<u>25.63</u>	64.80	44.92	<u>48.37</u>
	SpecPC	51.23	41.40	<b>23.37</b>	<b>62.26</b>	38.23	<b>43.66</b>	<b>56.84</b>	<u>44.48</u>	<b>25.91</b>	<b>67.37</b>	<b>47.15</b>	<b>48.44</b>

## 5.3 EFFICIENCY

We evaluate latency and memory on a single NVIDIA H200 (141GB) GPU. The target model is Qwen2.5-32B, with draft models of Qwen2.5-1.5B for KV dropping and Qwen2.5-0.5B for prompt compression. Latency is measured as the time to generate 64 tokens including all draft stages, with  $n_{\text{lookahead}}$  set to 64 for SpecKV and SpecKV-PC<sup>4</sup>, 8 for SpecPrefill and LAQ++, and 1 for SpecPC, with SpecKV-PC prompts compressed to 2048 tokens. For KV dropping methods, we report peak system memory, while for prompt compression, we report memory for the compression stage only,

<sup>4</sup>While some tasks can generate up to 128 tokens, benchmark outputs are typically under 64 tokens.

because the target model uses the same amount of memory for all prompt compression algorithms. A detailed breakdown of latency and memory is included in Section E.7.

SpecKV outperforms both SnapKV and LAQ++ in speed, driven by efficient sparse prefilling and a low-latency draft model (Fig. 6). Notably, SpecKV-PC achieves a 75% reduction in latency compared to LAQ++ at 64k context, as it significantly reduces the target prefill bottleneck through prompt compression. Regarding memory, while SpecKV requires more memory than SnapKV to store draft weights, this overhead is constant and can be further reduced by offloading to a CPU. Crucially, SpecKV is far more memory-efficient than LAQ++, which needs the entire target KV cache to function, matching the memory footprint of the dense target model. Thus, SpecKV offers a superior combination of accuracy and memory efficiency. Furthermore, SpecKV-PC yields substantial peak memory savings (around 25GB compared to LAQ++ at 64k context) by feeding a shorter prompt to the target model. Other KV dropping methods are omitted as their performance is similar to SnapKV.

SpecPC achieves the lowest latency among all baselines. It avoids the CPU preprocessing overhead of CPC and R2C and is faster than SpecPrefill due to a shorter lookahead. SpecPC is also the most memory-efficient, using substantially less memory than R2C. Overall, prompt compression is faster than KV dropping because only the compressed prompt tokens ( $C_{\max}$ ) are passed to the target model.

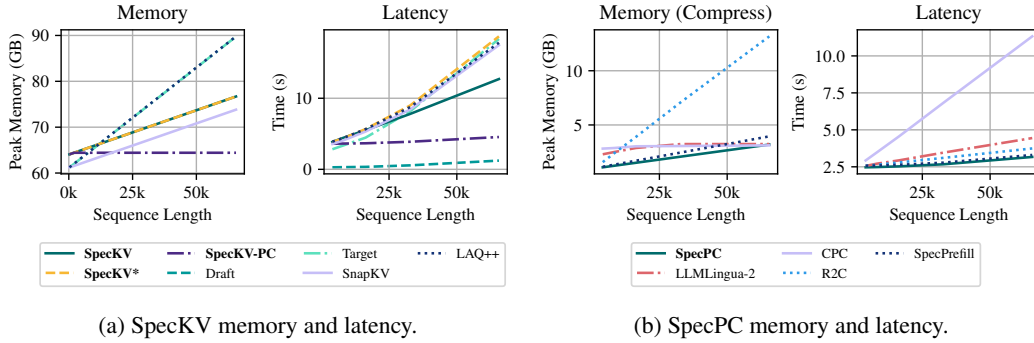


Figure 6: Peak memory usage and latency. SpecKV\* denotes SpecKV without sparse prefill.

## 6 DISCUSSION

In this paper, we present Draft-based Approximate Inference, a framework that leverages draft model lookahead for approximate inference. Within this framework, we introduce two concrete algorithms: Speculative KV Dropping (SpecKV) for KV cache dropping with sparse prefill, and Speculative Prompt Compression (SpecPC) for prompt compression. We further propose SpecKV-PC, a cascaded pipeline that synergizes these approaches for improved efficiency and accuracy. Our approach is grounded in theoretical and empirical analyses that justify lookahead-based KV cache dropping and the use of draft model token importance to approximate target model importance. Through comprehensive experiments on long-context benchmarks, we show that our methods consistently achieve state-of-the-art accuracy under fixed KV cache and prompt size constraints, surpassing prior baselines. These contributions establish draft model lookahead as a powerful tool for efficient long-context inference, extending the role of draft models beyond speculative decoding.

**Limitations and Future Work** For SpecKV, draft generation causes minimal latency even for larger  $n_{\text{lookahead}}$  (Section E.7). However, very long outputs or large  $n_{\text{lookahead}}$  values may reduce performance. In these cases, lowering  $n_{\text{lookahead}}$  could maintain speed with little loss in accuracy. For SpecPC, increasing  $n_{\text{lookahead}}$  to generate more tokens led to only minor accuracy gains; better leveraging longer drafts remains future work. Additionally, while our methods are robust using very small (Fig. 18) or cross-family (Section E.5) draft models, a reasonably accurate draft model is still required. Currently, Draft-based Approximate Inference supports sparse prefill, KV dropping, and prompt compression. Extensions such as lookahead-based sparse decoding or iterative KV cache dropping, where KV entries are periodically removed using draft lookahead, could further improve support for reasoning models with long outputs.

## REPRODUCIBILITY STATEMENT

We provide a link to our public code. All algorithms, datasets, and experimental details, including hyperparameter settings, can be found in Section 5 and the appendix (Sections A, C and D).

## REFERENCES

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint*, abs/2303.08774, 2023. URL <https://arxiv.org/abs/2303.08774>.
- Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. GPT-OSS-120B & GPT-OSS-20B model card. *arXiv preprint*, abs/2508.10925, 2025. URL <https://arxiv.org/abs/2508.10925>.
- Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. GQA: Training generalized multi-query Transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4895–4901, 2023.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibong Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint*, abs/2502.13923, 2025. URL <https://arxiv.org/abs/2502.13923>.
- 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. 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)*, pp. 3119–3137, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.172. URL <https://aclanthology.org/2024.acl-long.172>.
- Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document Transformer. *arXiv preprint*, abs/2004.05150, 2020. URL <https://arxiv.org/abs/2004.05150>.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64a-Abstract.html>.
- Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D Lee, Deming Chen, and Tri Dao. Medusa: Simple LLM inference acceleration framework with multiple decoding heads. In *International Conference on Machine Learning*, pp. 5209–5235. PMLR, 2024a.
- Zefan Cai, Yichi Zhang, Bofei Gao, Yuliang Liu, Tianyu Liu, Keming Lu, Wayne Xiong, Yue Dong, Baobao Chang, Junjie Hu, and Xiao Wen. PyramidKV: Dynamic KV cache compression based on pyramidal information funneling. *arXiv preprint*, abs/2406.02069, 2024b. URL <https://arxiv.org/abs/2406.02069>.
- E.J. Candes and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005. doi: 10.1109/TIT.2005.858979.
- Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. Accelerating large language model decoding with speculative sampling. *arXiv preprint*, abs/2302.01318, 2023. URL <https://arxiv.org/abs/2302.01318>.

- Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024.
- Eunseong Choi, Sunkyung Lee, Minjin Choi, Jun Park, and Jongwuk Lee. From reading to compressing: Exploring the multi-document reader for prompt compression. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 14734–14754, 2024.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Un-supervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL <https://aclanthology.org/2020.acl-main.747>.
- Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations*, 2024.
- Song Dingjie, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. MileBench: Benchmarking MLLMs in long context. In *First Conference on Language Modeling*, 2024.
- Xueying Du, Mingwei Liu, Kaixin Wang, Hanlin Wang, Junwei Liu, Yixuan Chen, Jiayi Feng, Chaofeng Sha, Xin Peng, and Yiling Lou. Evaluating large language models in class-level code generation. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering, ICSE ’24*, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400702174. doi: 10.1145/3597503.3639219. URL <https://doi.org/10.1145/3597503.3639219>.
- Yuan Feng, Junlin Lv, Yukun Cao, Xike Xie, and S Kevin Zhou. Ada-KV: Optimizing KV cache eviction by adaptive budget allocation for efficient LLM inference. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL <https://openreview.net/forum?id=tcisuhGsQZ>.
- Yu Fu, Zefan Cai, Abedelkadir Asi, Wayne Xiong, Yue Dong, and Wen Xiao. Not all heads matter: A head-level KV cache compression method with integrated retrieval and reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=FJFVmeXusW>.
- Gemma Team. Gemma 3. 2025. URL <https://google.com/Gemma3Report>.
- Google DeepMind. Gemini 2.5 Pro: Advanced reasoning AI model, 2025. URL <https://deepmind.google/technologies/gemini/pro/>.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint*, abs/2407.21783, 2024. URL <https://arxiv.org/abs/2407.21783>.
- Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. RULER: What’s the real context size of your long-context language models? In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=kIoBbc76Sy>.
- Yunhai Hu, Zining Liu, Zhenyuan Dong, Tianfan Peng, Bradley McDanel, and Sai Qian Zhang. Speculative decoding and beyond: An in-depth survey of techniques. *arXiv preprint*, abs/2502.19732, 2025. URL <https://arxiv.org/abs/2502.19732>.
- Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 874–880, Online, 2021. Association for Computational Linguistics. URL <https://aclanthology.org/2021.eacl-main.74>.

- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. Mistral 7B. *arXiv preprint*, abs/2310.06825, 2023a. URL <https://arxiv.org/abs/2310.06825>.
- Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LLMingua: Compressing prompts for accelerated inference of large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023b. URL <https://openreview.net/forum?id=ADsEdyI32n>.
- Huiqiang Jiang, YUCHENG LI, Chengruidong Zhang, Qianhui Wu, Xufang Luo, Surin Ahn, Zhenhua Han, Amir H. Abdi, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. MInference 1.0: Accelerating pre-filling for long-context LLMs via dynamic sparse attention. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a. URL <https://openreview.net/forum?id=fPBACAbqSN>.
- Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LongLLMingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1658–1677, 2024b.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.
- Xunhao Lai, Jianqiao Lu, Yao Luo, Yiyuan Ma, and Xun Zhou. FlexPrefill: A context-aware sparse attention mechanism for efficient long-sequence inference. In *The Thirteenth International Conference on Learning Representations*, 2025.
- Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from Transformers via speculative decoding. In *International Conference on Machine Learning*, pp. 19274–19286. PMLR, 2023.
- Yucheng Li, Bo Dong, Frank Guerin, and Chenghua Lin. Compressing context to enhance inference efficiency of large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL <https://openreview.net/forum?id=cjbdRN8Yxy>.
- Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. SnapKV: LLM knows what you are looking for before generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=poE54GOq2l>.
- Barys Liskavets, Maxim Ushakov, Shuvendu Roy, Mark Klibanov, Ali Etemad, and Shane K Luke. Prompt compression with context-aware sentence encoding for fast and improved LLM inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 24595–24604, 2025.
- Di Liu, Meng Chen, Baotong Lu, Huiqiang Jiang, Zhenhua Han, Qianxi Zhang, Qi Chen, Chengruidong Zhang, Bailu Ding, Kai Zhang, Chen Chen, Fan Yang, Yuqing Yang, and Lili Qiu. RetrievalAttention: Accelerating long-context LLM inference via vector retrieval. *arXiv preprint*, abs/2409.10516, 2024a. URL <https://arxiv.org/abs/2409.10516>.
- Fei Liu et al. Learning to summarize from human feedback. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 583–592, 2020.
- Jingyu Liu, Beidi Chen, and Ce Zhang. Speculative prefill: Turbocharging TTFT with lightweight and training-free token importance estimation. *arXiv preprint*, abs/2502.02789, 2025. URL <https://arxiv.org/abs/2502.02789>.
- Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024b.

- Piotr Nawrot, Robert Li, Renjie Huang, Sebastian Ruder, Kelly Marchisio, and Edoardo M. Ponti. The sparse frontier: Sparse attention trade-offs in Transformer LLMs. *arXiv preprint*, abs/2504.17768, 2025. URL <https://arxiv.org/abs/2504.17768>.
- Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Menglin Xia, Xufang Luo, Jue Zhang, Qingwei Lin, Victor Rühle, Yuqing Yang, Chin-Yew Lin, et al. LLMlingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 963–981, 2024.
- Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NIPS-W*, 2017.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text Transformer. *Journal of machine learning research*, 21(140):1–67, 2020.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL <https://aclanthology.org/D16-1264>.
- Ranajoy Sadhukhan, Jian Chen, Zhuoming Chen, Vashisth Tiwari, Ruihang Lai, Jinyuan Shi, Ian En-Hsu Yen, Avner May, Tianqi Chen, and Beidi Chen. MagicDec: Breaking the latency-throughput tradeoff for long context generation with speculative decoding. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=CS2JWaziYr>.
- Hanshi Sun, Zhuoming Chen, Xinyu Yang, Yuandong Tian, and Beidi Chen. TriForce: Lossless acceleration of long sequence generation with hierarchical speculative decoding. In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=HVK6nl3i97>.
- Jiaming Tang, Yilong Zhao, Kan Zhu, Guangxuan Xiao, Baris Kasikci, and Song Han. QUEST: Query-aware sparsity for efficient long-context LLM inference. In *Forty-first International Conference on Machine Learning*, 2024. URL <https://openreview.net/forum?id=KzACYw0MTV>.
- Rishabh Tiwari, Haocheng Xi, Aditya Tomar, Coleman Hooper, Sehoon Kim, Maxwell Horton, Mahyar Najibi, Michael W. Mahoney, Kurt Keutzer, and Amir Gholami. QuantSpec: Self-speculative decoding with hierarchical quantized KV cache. *arXiv preprint*, abs/2502.10424, 2025. URL <https://arxiv.org/abs/2502.10424>.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017. URL <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>.
- Yixuan Wang, Shiyu Ji, Yijun Liu, Yuzhuang Xu, Yang Xu, Qingfu Zhu, and Wanxiang Che. Lookahead Q-cache: Achieving more consistent KV cache eviction via pseudo query. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 34158–34174, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1732. URL <https://aclanthology.org/2025.emnlp-main.1732/>.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. HuggingFace’s Transformers: State-of-the-art natural language processing. *arXiv preprint*, abs/1910.03771, 2019. URL <https://arxiv.org/abs/1910.03771>.

- Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=NG7sS51zVF>.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint*, abs/2412.15115, 2024. URL <https://arxiv.org/abs/2412.15115>.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.
- An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, et al. Qwen2.5-1M technical report. *arXiv preprint*, abs/2501.15383, 2025b. URL <https://arxiv.org/abs/2501.15383>.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL <https://aclanthology.org/D18-1259>.
- Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Re, Clark Barrett, Zhangyang Wang, and Beidi Chen. H2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=RkRrPp7GKO>.



# Appendix

<b>A</b>	<b>Algorithm Pseudocode</b>	<b>17</b>
<b>B</b>	<b>Mathematical Proofs</b>	<b>17</b>
B.1	Proof of Theorem 1 . . . . .	19
B.2	Proof of Theorem 2 . . . . .	19
B.3	Proof of Theorem 3 . . . . .	20
<b>C</b>	<b>Benchmark Dataset Details</b>	<b>23</b>
C.1	LongBench . . . . .	23
C.2	RULER . . . . .	23
C.3	MileBench . . . . .	24
<b>D</b>	<b>Experimental Setup</b>	<b>24</b>
D.1	Hyperparameter Settings . . . . .	24
D.2	Implementation Details . . . . .	25
<b>E</b>	<b>Extended Experimental Analysis</b>	<b>26</b>
E.1	Additional Results on RULER and LongBench . . . . .	26
E.2	Integration and Comparison with the AdaKV Baseline . . . . .	26
E.3	Multi-modal Evaluation . . . . .	26
E.4	SpecKV: Impact of $n_{\text{lookahead}}$ Parameter on Importance Score Correlation . . . . .	27
E.5	Performance of Cross-Family Models . . . . .	27
E.6	Extended Results for SpecKV-PC . . . . .	28
E.7	SpecKV: Extended Latency and Memory Analysis . . . . .	29
E.8	Ablation Studies . . . . .	29
E.8.1	SpecKV: Analysis of Enhanced Draft Models . . . . .	30
E.8.2	SpecKV and SpecPC: Impact of $n_{\text{lookahead}}$ . . . . .	30
E.8.3	SpecKV: Accuracy Analysis of Sparse Prefill . . . . .	30

## A ALGORITHM PSEUDOCODE

---

### Algorithm 1 SpecKV

---

```

1: Input:
   Input sequence  $x$  with length  $n_{\text{in}}$ 
   Parameters: Number of lookahead tokens  $n_{\text{lookahead}}$ , Maximum cache capacity  $C_{\text{max}}$ ,
   Compression window size  $n_{\text{window}}$ , Kernel size  $k$ , Prefill window size  $n_{\text{slash}}$ ,
   Number of global tokens in prefill  $n_{\text{vert}}$ 

2: Generate a draft output  $y_{\text{draft}}$  of length  $n_{\text{lookahead}}$  using the draft model.
3: Forward  $x$  and  $y_{\text{draft}}$  through the target model.
4: for each attention head in target model do
5:    $X \leftarrow$  Target model hidden states from prompt at current layer  $\triangleright X \in \mathbb{R}^{n_{\text{in}} \times d}$ 
6:    $Y \leftarrow$  Target model hidden states from draft output at current layer  $\triangleright Y \in \mathbb{R}^{n_{\text{lookahead}} \times d}$ 
7:    $m \leftarrow n_{\text{in}} - n_{\text{window}}$ 
8:    $A \leftarrow$  CrossAttention ( $Q$  from  $[X_{Y^m}^m]$ ,  $K/V$  from  $X_{:m}$ )  $\triangleright$  Compute attention
9:    $s \leftarrow \text{MaxReduce}(A)$   $\triangleright \mathbb{R}^{n_1 \times n_2} \rightarrow \mathbb{R}^{n_2}$ 
10:   $s \leftarrow \text{AvgPool1D}(s, k)$   $\triangleright$  Smooth attention
11:   $i_{\text{vert}} \leftarrow \text{topk}(s, n_{\text{vert}})$   $\triangleright$  Select global tokens (Sparse prefill)
12:   $i_{\text{slash}} \leftarrow \{1, 2, \dots, n_{\text{slash}}\}$   $\triangleright$  Sliding window (Sparse prefill)
13:   $i_{\text{cache}} \leftarrow \text{topk}(s, C_{\text{max}} - n_{\text{window}}) \cup \{m + 1, m + 2, \dots, n_{\text{in}}\}$   $\triangleright$  Select KVs
14:  output  $\leftarrow \text{VerticalSlash}(X, i_{\text{vert}}, i_{\text{slash}})$   $\triangleright$  Sparse prefill attention
15:  cache  $\leftarrow K_{i_{\text{cache}}}, V_{i_{\text{cache}}}$   $\triangleright$  KV cache dropping
16: end for

```

---



---

### Algorithm 2 SpecPC

---

```

1: Input:
   Draft attention tensor  $A \in \mathbb{R}^{n_{\text{layer}} \times n_{\text{head}} \times (n_{\text{in}} + n_{\text{lookahead}} - 1) \times n_{\text{in}}}$ 
   Parameters: Window size  $n_{\text{window}}$ , Kernel size  $k$ , Number of neighbors  $n_{\text{neighbor}}$ ,
   Number of selected tokens  $C_{\text{max}}$ , Number of skipped layers  $l_{\text{skip}}$ 

2:  $m \leftarrow n_{\text{in}} - n_{\text{window}}$ 
3:  $A \leftarrow A_{l_{\text{skip}}:, :, m:, :m}$   $\triangleright$  Skip layers and only consider window queries and non-window keys
4: for  $j \in \{1, 2, \dots, n_{\text{window}}\}$  do
5:    $A_{\dots, j, :} \leftarrow \frac{j}{n_{\text{window}}} A_{\dots, j, :}$   $\triangleright$  Assign more weight to later tokens
6: end for
7:  $s \leftarrow \text{MaxReduce}(A)$   $\triangleright \mathbb{R}^{n_1 \times n_2 \times n_3 \times n_4} \rightarrow \mathbb{R}^{n_4}$ 
8:  $s \leftarrow \text{AvgPool1D}(s, k)$   $\triangleright$  Smooth attention
9:  $s \leftarrow \text{MaxPool1D}(s, n_{\text{neighbor}})$   $\triangleright$  Keep neighbor tokens
10:  $i_{\text{selected}} \leftarrow \text{topk}(s, C_{\text{max}}) \cup \{m + 1, m + 2, \dots, n_{\text{in}}\}$   $\triangleright$  Keep most activated tokens and window tokens
11: return  $i_{\text{selected}}$ 

```

---

## B MATHEMATICAL PROOFS

**Lemma 1.**  $\|\text{Softmax}(x) - \text{Softmax}(y)\|_2 \leq \|x - y\|_\infty$ .

*Proof.* Let  $J$  be the Jacobian matrix of Softmax. Then,  $J(v) = \text{diag}(p) - pp^T$ , where  $p = \text{Softmax}(v)$ . Note that  $p$  is a probability distribution, so  $p_i \geq 0$  for all  $i$  and  $\sum_i p_i = 1$ . For any

vectors  $v$  and  $z$ ,

$$\begin{aligned}
\|J(v)z\|_2^2 &= \|(\text{diag}(p) - pp^T)z\|_2^2 \\
&= \sum_i (p_i z_i - p_i p^T z)^2 \\
&= \sum_i p_i^2 (z_i - p^T z)^2 \\
&\leq \sum_i p_i (z_i - p^T z)^2 \\
&= \sum_i (p_i z_i^2 - 2p_i z_i p^T z + p_i (p^T z)^2) \\
&= \sum_i p_i z_i^2 - (p^T z)^2 \\
&\leq \sum_i p_i z_i^2 \\
&\leq \sum_i p_i \|z\|_\infty^2 \\
&\leq \|z\|_\infty^2.
\end{aligned}$$

Thus,  $\|J(v)z\|_2 \leq \|z\|_\infty$  for all  $v$  and  $z$ . From the fundamental theorem of line integrals,

$$\text{Softmax}(x) - \text{Softmax}(y) = \int_0^1 J(y + t(x - y))(x - y) dt. \quad (4)$$

Finally,

$$\begin{aligned}
\|\text{Softmax}(x) - \text{Softmax}(y)\|_2 &= \left\| \int_0^1 J(y + t(x - y))(x - y) dt \right\|_2 \\
&\leq \int_0^1 \|J(y + t(x - y))(x - y)\|_2 dt \\
&\leq \int_0^1 \|x - y\|_\infty dt \\
&= \|x - y\|_\infty.
\end{aligned}$$

□

**Lemma 2.** Let  $y = \text{Softmax}(x)$  and  $y' = \text{Softmax}(x')$ . If  $\|y - y'\|_p \leq \epsilon$ , then there exists a scalar  $c$  such that  $\|x - x' - c\mathbf{1}\|_p \leq \frac{\epsilon}{m}$ , where  $m = \min_i (\min(y_i, y'_i))$  and  $p \in \{1, 2, \dots, \infty\}$ .

*Proof.* From the mean value theorem, there exists  $\xi \in (y_i, y'_i)$  such that

$$\frac{\log y_i - \log y'_i}{y_i - y'_i} = \frac{d \log t}{dt} \bigg|_{t=\xi} = \frac{1}{\xi}. \quad (5)$$

Note that  $\xi > 0$ . Then,

$$|\log y_i - \log y'_i| = \frac{1}{\xi} |y_i - y'_i| \leq \frac{1}{m} |y_i - y'_i|. \quad (6)$$

Let  $c = \log \sum_j e^{x_j} - \log \sum_j e^{x'_j}$ , so

$$\left| \log \frac{e^{x_i}}{\sum_j e^{x_j}} - \log \frac{e^{x'_i}}{\sum_j e^{x'_j}} \right| = \left| x_i - x'_i - \left( \log \sum_j e^{x_j} - \log \sum_j e^{x'_j} \right) \right| = |x_i - x'_i - c| \quad (7)$$

for all  $i$ . Thus,

$$\begin{aligned}
|x_i - x'_i - c| &\leq \frac{1}{m} |y_i - y'_i| \implies |x_i - x'_i - c|^p \leq \frac{1}{m^p} |y_i - y'_i|^p \\
&\implies \sum_i |x_i - x'_i - c|^p \leq \frac{1}{m^p} \sum_i |y_i - y'_i|^p \\
&\implies \|x - x' - c\mathbf{1}\|_p^p \leq \frac{1}{m^p} \|y - y'\|_p^p \\
&\implies \|x - x' - c\mathbf{1}\|_p \leq \frac{1}{m} \|y - y'\|_p \\
&\implies \|x - x' - c\mathbf{1}\|_p \leq \frac{\epsilon}{m}.
\end{aligned}$$

□

### B.1 PROOF OF THEOREM 1

We define the vector of importance scores as and its approximation as

$$s^T = \frac{1}{n_{\text{out}}} \sum_{i=1}^{n_{\text{out}}} \text{Softmax} \left( \frac{x_i^{(o)T} W_q W_k^T X^T}{\sqrt{d}} \right) \quad \text{and} \quad \hat{s}^T = \frac{1}{n_{\text{out}}} \sum_{i=1}^{n_{\text{out}}} \text{Softmax} \left( \frac{\hat{x}_i^{(o)T} W_q W_k^T X^T}{\sqrt{d}} \right), \quad (8)$$

where  $X = [x_1, \dots, x_{n_{\text{in}}}]^T \in \mathbb{R}^{n_{\text{in}} \times d}$  is the matrix of input embeddings,  $x_i^{(o)} \in \mathbb{R}^d$  is the  $i$ th output embedding, and  $\hat{x}_i^{(o)} \in \mathbb{R}^d$  is the  $i$ th approximate output embedding (from the draft model).  $s_i$  and  $\hat{s}_i$  denote the importance of the  $i$ th KV pair. In practice, SpecKV estimates importance using queries from recent input and draft output tokens. This is omitted from the theoretical analysis for clarity.

*Proof.* We assume  $\|x_i^{(o)} - \hat{x}_i^{(o)}\|_2 \leq \epsilon$  for all  $i$  and  $\|x_j\|_2 \leq \sqrt{d}$  for all  $j$ .

$\|x_i^{(o)} - \hat{x}_i^{(o)}\|_2 \leq \epsilon$ , so

$$\begin{aligned}
\left| \frac{x_i^{(o)T} W_q W_k^T x_j}{\sqrt{d}} - \frac{\hat{x}_i^{(o)T} W_q W_k^T x_j}{\sqrt{d}} \right| &= \frac{1}{\sqrt{d}} \left| (x_i^{(o)} - \hat{x}_i^{(o)})^T W_q W_k^T x_j \right| \\
&\leq \frac{1}{\sqrt{d}} \|W_q W_k^T\|_2 \epsilon \sqrt{d} \\
&= \epsilon \|W_q W_k^T\|_2.
\end{aligned}$$

Thus,

$$\left\| \frac{x_i^{(o)T} W_q W_k^T X^T}{\sqrt{d}} - \frac{\hat{x}_i^{(o)T} W_q W_k^T X^T}{\sqrt{d}} \right\|_{\infty} \leq \epsilon \|W_q W_k^T\|_2. \quad (9)$$

Applying Lemma 1 and the triangle inequality, we get

$$\begin{aligned}
\|s - \hat{s}\|_2 &= \left\| \frac{1}{n_{\text{out}}} \sum_{i=1}^{n_{\text{out}}} \text{Softmax} \left( \frac{x_i^{(o)T} W_q W_k^T X^T}{\sqrt{d}} \right) - \frac{1}{n_{\text{out}}} \sum_{i=1}^{n_{\text{out}}} \text{Softmax} \left( \frac{\hat{x}_i^{(o)T} W_q W_k^T X^T}{\sqrt{d}} \right) \right\|_2 \\
&\leq \frac{1}{n_{\text{out}}} \sum_{i=1}^{n_{\text{out}}} \left\| \text{Softmax} \left( \frac{x_i^{(o)T} W_q W_k^T X^T}{\sqrt{d}} \right) - \text{Softmax} \left( \frac{\hat{x}_i^{(o)T} W_q W_k^T X^T}{\sqrt{d}} \right) \right\|_2 \\
&\leq \frac{1}{n_{\text{out}}} \sum_{i=1}^{n_{\text{out}}} \left\| \frac{x_i^{(o)T} W_q W_k^T X^T}{\sqrt{d}} - \frac{\hat{x}_i^{(o)T} W_q W_k^T X^T}{\sqrt{d}} \right\|_{\infty} \\
&\leq \epsilon \|W_q W_k^T\|_2,
\end{aligned}$$

□

### B.2 PROOF OF THEOREM 2

$$X = [x_1, \dots, x_{n_{\text{in}}}]^T \quad (10)$$

$$Y = [y_1, \dots, y_{n_{\text{in}}}]^T = \text{Softmax} \left( \frac{X W_q W_k^T X^T}{\sqrt{d}} \right) X W_v \quad (11)$$

$$\hat{Y} = [\hat{y}_1, \dots, \hat{y}_{n_{in}}]^T = \text{Softmax} \left( \frac{X \hat{W}_q \hat{W}_k^T X^T}{\sqrt{d}} \right) X \hat{W}_v \quad (12)$$

$$A = [a_1, \dots, a_{n_{in}}]^T = \text{Softmax} \left( \frac{X W_q W_k^T X^T}{\sqrt{d}} \right) \quad (13)$$

$$\hat{A} = [\hat{a}_1, \dots, \hat{a}_{n_{in}}]^T = \text{Softmax} \left( \frac{X \hat{W}_q \hat{W}_k^T X^T}{\sqrt{d}} \right) \quad (14)$$

*Proof.* We assume  $\|y_i - \hat{y}_i\|_2 \leq \epsilon \|X\|_{\infty,2}$ , where  $\|X\|_{\infty,2}$  is the maximum  $\ell_2$  norm of the rows of  $X$ . Additionally, we assume that there exists a constant  $c$  such that  $cX^T$  has the Restricted Isometry Property (Candes & Tao, 2005) with parameters  $(2k, \delta)$ , where  $\delta$  is the restricted isometry constant and  $k$  is the approximate sparsity of  $a_i$  and  $\hat{a}_i$ .

Recall that a matrix  $B$  satisfies the Restricted Isometry Property with constant  $\delta \in (0, 1)$  if for every  $k$ -sparse vector  $v$ , the following inequality holds:

$$(1 - \delta)\|v\|_2^2 \leq \|Bv\|_2^2 \leq (1 + \delta)\|v\|_2^2. \quad (15)$$

Let  $\Delta W_v = W_v - \hat{W}_v$  and  $\Delta a_i = a_i - \hat{a}_i$ .

If  $n_{in} = 1$ , then  $A = \hat{A} \in \mathbb{R}^{1 \times 1}$  with  $A_{1,1} = \hat{A}_{1,1} = 1$ , so  $AX = \hat{A}X = X = x_1^T$ , which implies

$$\|y_1 - \hat{y}_1\|_2 = \|x_1^T W_v - x_1^T \hat{W}_v\|_2 = \|x_1^T \Delta W_v\|_2 \leq \epsilon \|X\|_{\infty,2} = \epsilon \|x_1\|_2 \quad (16)$$

for all  $x_1$ .  $\|x_1^T W_v\|_2 \leq \epsilon \|x_1\|_2$  for all  $x_1$  is the definition of the matrix  $\ell_2$  norm, so  $\|\Delta W_v\|_2 \leq \epsilon$ .

$$\begin{aligned} \|y_i - \hat{y}_i\|_2 &= \|a_i^T X W_v - \hat{a}_i^T X \hat{W}_v\|_2 \\ &= \|a_i^T X W_v - (\hat{a}_i^T X W_v - \hat{a}_i^T X \Delta W_v)\|_2 \\ &= \|a_i^T X W_v - (a_i^T X W_v - \Delta a_i^T X W_v - \hat{a}_i^T X \Delta W_v)\|_2 \\ &= \|\Delta a_i^T X W_v + \hat{a}_i^T X \Delta W_v\|_2 \\ &\leq \epsilon \|X\|_{\infty,2} \end{aligned}$$

Then,  $\|\Delta a_i^T X W_v\|_2 \leq \epsilon \|X\|_{\infty,2} + \|\hat{a}_i^T X \Delta W_v\|_2$ .

Since  $\hat{a}_i^T X$  is a convex combination of the rows of  $X$ ,  $\|\hat{a}_i^T X\|_2 \leq \|X\|_{\infty,2}$ .

Thus,  $\|\Delta a_i^T X W_v\|_2 \leq \epsilon \|X\|_{\infty,2} + \|\hat{a}_i^T X \Delta W_v\|_2 \leq \epsilon \|X\|_{\infty,2} + \|\hat{a}_i^T X\|_2 \|\Delta W_v\|_2 \leq 2\epsilon \|X\|_{\infty,2}$ .

Attention scores are approximately sparse (Jiang et al., 2024a), especially for long sequences. Therefore, we assume  $a_i$  and  $\hat{a}_i$  are  $k$ -sparse. Then,  $\Delta a_i$  is at most  $2k$ -sparse. Since  $cX^T$  has the Restricted Isometry Property with parameters  $2k, \delta$ ,

$$(1 - \delta)\|\Delta a_i\|_2 \leq \|\Delta a_i^T (cX)\|_2 \leq (1 + \delta)\|\Delta a_i\|_2. \quad (17)$$

Then,

$$\frac{1}{c}(1 - \delta)\|\Delta a_i\|_2 \leq \|\Delta a_i^T X\|_2 \leq \frac{2\epsilon \|X\|_{\infty,2}}{\sigma_{\min}(W_v)}, \quad (18)$$

so

$$\|\Delta a_i\|_2 \leq \frac{2c\epsilon \|X\|_{\infty,2}}{\sigma_{\min}(W_v)(1 - \delta)}. \quad (19)$$

□

### B.3 PROOF OF THEOREM 3

**Theorem 3.** If  $\|Y - \hat{Y}\|_2 \leq \epsilon \|X\|_2$  for all  $X$  and the column space of  $W_q, W_k, \hat{W}_q, \hat{W}_k$  is a subset of the column space of  $W_v$ , then  $\|a_i - \hat{a}_i\|_2 \leq \epsilon \delta$ , where

$$\delta = 2d \frac{\sigma_{\max}(W_v)^2}{\sigma_{\min}(W_v)} \exp \left( 2 \max \left( \frac{\|W_q\|_2 \|W_k\|_2}{\sigma_{\min}(W_v)^2}, \frac{\|\hat{W}_q\|_2 \|\hat{W}_k\|_2}{\sigma_{\min}(\hat{W}_v)^2} \right) \right) \|X\|_{\infty,2}^2. \quad (20)$$

*Proof.* We assume  $\|Y - \hat{Y}\|_2 \leq \epsilon\|X\|_2$ . Additionally, we assume that the column space of  $W_q, W_k, \hat{W}_q, \hat{W}_k$  is a subset of the column space of  $W_v$ . To get a norm bound on  $\Delta a_i = a_i - \hat{a}_i$ , we will bound the norms of the error in approximate weight matrices. We will find these bounds by using specific inputs, taking advantage of the fact that  $\|Y - \hat{Y}\|_2 \leq \epsilon\|X\|_2$  for all  $X$ .

We will start by bounding  $\Delta W_v = W_v - \hat{W}_v$ , by choosing an input that fixes  $A$  and  $\hat{A}$ . If  $n = 1$ , then  $A = \hat{A} = [1]$ , so  $AX = \hat{A}X = X = x_1^T$ , which implies

$$\|Y - \hat{Y}\|_2 = \|AXW_v - \hat{A}X\hat{W}_v\|_2 = \|x_1^T W_v - x_1^T \hat{W}_v\|_2 = \|x_1^T \Delta W_v\|_2 \leq \epsilon\|X\|_2 = \epsilon\|x_1\|_2 \quad (21)$$

for all  $x_1$ . Thus,  $\|\Delta W_v\|_2 \leq \epsilon$ .

Next, we will bound the norm of  $\Delta B = B - \hat{B}$ , where  $B = W_q W_k^T$  and  $\hat{B} = \hat{W}_q \hat{W}_k^T$ . We will choose the  $X$  so that the values are the identity matrix. Then  $Y = A$ . Let  $U\Sigma V^T$  be the singular value decomposition of  $W_v$ . We set

$$X = \Phi V \Sigma^{-1} U^T, \quad (22)$$

where  $\Phi$  is an arbitrary orthonormal basis spanning  $\mathbb{R}^{d \times d}$ .

Note that  $\sigma_{\min}(\Phi) = \sigma_{\max}(\Phi) = 1$ , so  $\|X\|_2 = \sigma_{\max}(X) = \frac{1}{\sigma_{\min}(W_v)}$  and  $\sigma_{\min}(X) = \frac{1}{\sigma_{\max}(W_v)}$ .

Now,

$$\begin{aligned} \|Y - \hat{Y}\|_2 &= \|AXW_v - \hat{A}X\hat{W}_v\|_2 \\ &= \|AXW_v - (\hat{A}XW_v - \hat{A}X\Delta W_v)\|_2 \\ &= \|AXW_v - (AXW_v - \Delta AXW_v - \hat{A}X\Delta W_v)\|_2 \\ &= \|\Delta AXW_v + \hat{A}X\Delta W_v\|_2 \\ &\leq \|\Delta AXW_v\|_2 + \|\hat{A}X\Delta W_v\|_2 \\ &= \|\Delta A\Phi V \Sigma^{-1} U^T U \Sigma V^T\|_2 + \|\hat{A}X\Delta W_v\|_2 \\ &= \|\Delta A\Phi\|_2 + \|\hat{A}X\Delta W_v\|_2 \\ &= \|\Delta A\|_2 + \|\hat{A}X\Delta W_v\|_2 \\ &\leq \epsilon\|X\|_2 + \epsilon\|X\|_2 \\ &\leq \frac{2\epsilon}{\sigma_{\min}(W_v)}. \end{aligned}$$

Note that each row of  $\hat{A}$  is a probability distribution (non-negative entries summing to 1), so left-multiplying  $X$  by  $\hat{A}$  forms a convex combination of the rows of  $X$ . From Jensen's inequality we get  $\|\hat{A}X\|_2 \leq \|X\|_2$ , because  $x \rightarrow \|x\|_2$  is a convex function.

Let  $\delta_1 = \max\left(\frac{\|W_q\|_2\|W_k\|_2}{\sigma_{\min}(W_v)^2}, \frac{\|\hat{W}_q\|_2\|\hat{W}_k\|_2}{\sigma_{\min}(\hat{W}_v)^2}\right)$ . Since  $\|X\|_2 = \frac{1}{\sigma_{\min}(W_v)}$  and  $\|B\|_2 \leq \|W_q\|_2\|W_k\|_2$ ,  $\|XB X^T\|_2 \leq \delta_1$ . Consequently,  $|x_i^T B x_j| \leq \delta_1$  for all  $i, j$ . This implies that each attention weight satisfies

$$a_{i,j} > \frac{e^{-\delta_1}}{\sum_{j=1}^d e^{\delta_1}} = \frac{1}{d} e^{-2\delta_1}. \quad (23)$$

The same argument applied to  $\hat{a}$  gives

$$\hat{a}_{i,j} > \frac{e^{-\delta_1}}{\sum_{j=1}^d e^{\delta_1}} = \frac{1}{d} e^{-2\delta_1}. \quad (24)$$

Applying Lemma 2 to each row, there exists  $c \in \mathbb{R}^d$  such that

$$\begin{aligned}
\left\| \frac{XB X^T}{\sqrt{d}} - \frac{X\hat{B} X^T}{\sqrt{d}} + c\mathbf{1}^T \right\|_2 &\leq d e^{2\delta_1} \left\| \text{Softmax} \left( \frac{XB X^T}{\sqrt{d}} \right) - \text{Softmax} \left( \frac{X\hat{B} X^T}{\sqrt{d}} \right) \right\|_2 \\
\left\| XB X^T - X\hat{B} X^T + \sqrt{d}c\mathbf{1}^T \right\|_2 &\leq d^{3/2} e^{2\delta_1} \|\Delta A\|_2 \\
&\leq \frac{2\epsilon d^{3/2} e^{2\delta_1}}{\sigma_{\min}(W_v)}.
\end{aligned} \tag{25}$$

Minimizing over  $c$ , we obtain

$$\begin{aligned}
\min_c \|XB X^T - X\hat{B} X^T + \sqrt{d}c\mathbf{1}^T\|_2 &= \min_c \|X\Delta B X^T - \sqrt{d}c\mathbf{1}^T\|_2 \\
&= \|X\Delta B X^T - \frac{1}{d}X\Delta B X^T \mathbf{1}\mathbf{1}^T\|_2 \\
&= \|X\Delta B X^T (I - \frac{1}{d}\mathbf{1}\mathbf{1}^T)\|_2
\end{aligned} \tag{26}$$

Substituting in the definition of  $X$ , we get

$$\|\Phi V \Sigma^{-1} U^T \Delta B U \Sigma^{-1} V^T \Phi^T (I - d^{-1} \mathbf{1}\mathbf{1}^T)\|_2 \leq \frac{\epsilon d^{3/2} e^{2\delta_1}}{\sigma_{\min}(W_v)}. \tag{27}$$

Each multiplication by  $\Sigma^{-1}$  can decrease the norm by at most  $\frac{1}{\sigma_{\max}(W_v)}$ , so when removing both instances of  $\Sigma^{-1}$  we scale the bound by  $\sigma_{\max}(W_v)^2$ , giving us

$$\|\Phi V U^T \Delta B U V^T \Phi^T (I - d^{-1} \mathbf{1}\mathbf{1}^T)\|_2 \leq \frac{2\epsilon d^{3/2} e^{2\delta_1}}{\sigma_{\min}(W_v)} \sigma_{\max}(W_v)^2. \tag{28}$$

Then, since  $\Phi$  and  $V$  are orthonormal and preserve spectral norm under multiplication, we conclude

$$\|U^T \Delta B U V^T \Phi^T (I - d^{-1} \mathbf{1}\mathbf{1}^T)\|_2 \leq \frac{2\epsilon d^{3/2} e^{2\delta_1}}{\sigma_{\min}(W_v)} \sigma_{\max}(W_v)^2. \tag{29}$$

Finally, since the column space of  $W_q, W_k, \hat{W}_q, \hat{W}_k$  is a subset of the column space of  $W_v$ , the column space of  $\Delta B$  is a subset of the column space of  $U$ . Thus, left multiplication by  $U^T$  does not impact the spectral norm, so

$$\|\Delta B U V^T \Phi^T (I - d^{-1} \mathbf{1}\mathbf{1}^T)\|_2 \leq \frac{2\epsilon d^{3/2} e^{2\delta_1}}{\sigma_{\min}(W_v)} \sigma_{\max}(W_v)^2. \tag{30}$$

Note that the matrix  $I - \frac{1}{d}\mathbf{1}\mathbf{1}^T$  is a projection onto the subspace orthogonal to the all-ones vector. Its singular values are  $[1, \dots, 1, 0]$ , so its spectral norm is

$$\|I - \frac{1}{d}\mathbf{1}\mathbf{1}^T\|_2 = 1. \tag{31}$$

Moreover, since  $\|I - \frac{1}{h}\mathbf{1}\mathbf{1}^T\|_2 = 1$  and  $\Phi$  is an arbitrary orthonormal basis of  $\mathbb{R}^d$ , it follows that for any fixed  $P \in \mathbb{R}^{d \times d}$ , we can choose  $\Phi$  such that the largest component of  $P\Phi^T$  lies entirely in the subspace orthogonal to  $\mathbf{1}$ . In this case,

$$\|P\Phi^T (I - \frac{1}{d}\mathbf{1}\mathbf{1}^T)\|_2 = \|P\|_2. \tag{32}$$

Thus,  $\|\Delta B\|_2 \leq \delta_2$  where  $\delta_2 = 2\epsilon d^{3/2} \frac{\sigma_{\max}(W_v)^2}{\sigma_{\min}(W_v)} e^{2\delta_1}$ .

Now that we have bounded  $\|\Delta B\|_2$ , we will consider any input  $X$ . Then,  $|x_i^T \Delta B x_j| \leq \delta_2 \|X\|_{\infty,2}^2$ , so  $\|x_i^T \Delta B X^T\|_\infty \leq \delta_2 \|X\|_{\infty,2}^2$ .  $\|X\|_{\infty,2}$  is the maximum  $\ell_2$  norm of the rows of  $X$ .

From Lemma 1,

$$\begin{aligned}
\|a_i - \hat{a}_i\|_2 &= \left\| \text{Softmax} \left( \frac{x_i^T B X^T}{\sqrt{d}} \right) - \text{Softmax} \left( \frac{x_i^T \hat{B} X^T}{\sqrt{d}} \right) \right\|_2 \\
&\leq \left\| \frac{x_i^T B X^T}{\sqrt{d}} - \frac{x_i^T \hat{B} X^T}{\sqrt{d}} \right\|_\infty = \frac{1}{\sqrt{d}} \|x_i^T \Delta B X^T\|_\infty \leq \frac{\delta_2 \|X\|_{\infty,2}^2}{\sqrt{d}}.
\end{aligned}$$



$$\|a_i - \hat{a}_i\|_2 \leq \frac{\delta_2 \|X\|_{\infty,2}^2}{\sqrt{d}} = 2\epsilon d \frac{\sigma_{\max}(W_v)^2}{\sigma_{\min}(W_v)} \exp\left(2 \max\left(\frac{\|W_q\|_2 \|W_k\|_2}{\sigma_{\min}(W_v)^2}, \frac{\|\hat{W}_q\|_2 \|\hat{W}_k\|_2}{\sigma_{\min}(\hat{W}_v)^2}\right)\right) \|X\|_{\infty,2}^2. \quad (33)$$

□

## C BENCHMARK DATASET DETAILS

### C.1 LONGBENCH

Table 3: LongBench tasks.

Task	Dataset	Source	Avg. Words	Metric	Language	Size
<b>Single-Document QA</b>						
1-1	NarrativeQA	Literature, Film	18,409	F1	English	200
1-2	Qasper	Science	3,619	F1	English	200
1-3	MultiFieldQA-en	Multi-field	4,559	F1	English	150
1-4	MultiFieldQA-zh	Multi-field	6,701	F1	Chinese	200
<b>Multi-Document QA</b>						
2-1	HotpotQA	Wikipedia	9,151	F1	English	200
2-2	2WikiMultihopQA	Wikipedia	4,887	F1	English	200
2-3	MuSiQue	Wikipedia	11,214	F1	English	200
2-4	DuReader	Baidu Search	15,768	Rouge-L	Chinese	200
<b>Summarization</b>						
3-1	GovReport	Government report	8,734	Rouge-L	English	200
3-2	QMSum	Meeting	10,614	Rouge-L	English	200
3-3	MultiNews	News	2,113	Rouge-L	English	200
3-4	VCSUM	Meeting	15,380	Rouge-L	Chinese	200
<b>Few-shot Learning</b>						
4-1	TREC	Web question	5,177	Accuracy (CLS)	English	200
4-2	TriviaQA	Wikipedia, Web	8,209	F1	English	200
4-3	SAMSum	Dialogue	6,258	Rouge-L	English	200
4-4	LSHT	News	22,337	Accuracy (CLS)	Chinese	200
<b>Synthetic Task</b>						
5-1	PassageCount	Wikipedia	11,141	Accuracy (EM)	English	200
5-2	PassageRetrieval-en	Wikipedia	9,289	Accuracy (EM)	English	200
5-3	PassageRetrieval-zh	C4 Dataset	6,745	Accuracy (EM)	Chinese	200
<b>Code Completion</b>						
6-1	LCC	Github	1,235	Edit Sim	Python/C#/Java	500
6-2	RepoBench-P	Github repository	4,206	Edit Sim	Python/Java	500

LongBench<sup>5</sup> (Bai et al., 2024) is a benchmark suite designed for long-context evaluation, comprising 14 English tasks, five Chinese tasks, and two code tasks. As Llama does not support Chinese, we excluded the corresponding tasks. Furthermore, we removed the synthetic tasks, as these are already covered by the RULER benchmark. The remaining tasks are grouped into five categories: single-document question answering, multi-document question answering, summarization, few-shot learning, and code completion. For each category, the overall score is calculated as the average of all its subtasks. The final LongBench score is computed as the average across all included tasks. Table 3 provides an overview of all tasks, adapted from Bai et al. (2024).

### C.2 RULER

RULER<sup>6</sup> (Hsieh et al., 2024) is a synthetic dataset designed to evaluate the true supported context length of LLMs. It comprises 13 tasks, including eight needle-in-a-haystack (NIAH) retrieval tasks, two aggregation tasks, two question answering (QA) tasks, and one multi-hop tracing task.

The NIAH tasks involve hiding random key-value pairs within generated text and challenging the model to retrieve them. Aggregation tasks simulate summarization by asking the model to extract the

<sup>5</sup><https://huggingface.co/datasets/THUDM/LongBench> (MIT License)

<sup>6</sup><https://github.com/NVIDIA/RULER> (Apache License 2.0)

most frequent or common words from a given passage. The QA tasks require the model to answer a question about a randomly selected paragraph within the context, serving as a real-world analog to NIAH tasks. In the multi-hop tracing task, the model must identify all variable names that reference the same value within a chain of assignments.

RULER is generated for a range of sequence lengths using randomly generated texts drawn from Paul Graham essays, SQuAD (Rajpurkar et al., 2016), and HotPotQA (Yang et al., 2018) datasets. This approach enables a comprehensive assessment of a language model’s capability to process varying context lengths. Evaluation is conducted based on accuracy, considering a response correct if it contains the requested value associated with the specified key.

### C.3 MILEBENCH

Table 4: Overview of the MileBench datasets. Average tokens are computed using Qwen2.5-VL (Bai et al., 2025).

Category	Dataset	Avg. Words	Avg. Images	Avg. Tokens	Metric	Size
<b>Temporal</b>	EgocentricNavigation	85	45	3,079	Accuracy	200
	MovingDirection	62	5	1,042	Accuracy	200
	SceneTransition	66	20	5,125	Accuracy	200
<b>Semantic</b>	SlideVQA	66	2	2,053	Accuracy	200
	TQA	50	8	5,536	Accuracy	200
	WebQA	146	2	1,706	Accuracy	200

MileBench<sup>7</sup> (Dingjie et al., 2024) is a long-context benchmark designed to evaluate Multimodal Large Language Models (MLLMs). It comprises 29 multi-image-text datasets, organized into 12 tasks, which are further grouped into four categories: Temporal Multi-Image, Semantic Multi-Image, and two diagnostic categories—NIAH and Image Retrieval.

For our additional experiments in Section E.3, we selected three datasets each from the Temporal Multi-Image and Semantic Multi-Image categories: EgocentricNavigation, MovingDirection, and SceneTransition for the Temporal Multi-Image category, and SlideVQA, TQA, and WebQA for the Semantic Multi-Image category. Table 4 provides an overview of the selected datasets.

## D EXPERIMENTAL SETUP

### D.1 HYPERPARAMETER SETTINGS

Table 5: Prompt compression backbones and parameter counts.

Method	Backbone	Parameters
LLMLingua-2 (Pan et al., 2024)	xlm-roberta-large (Conneau et al., 2020)	560M
CPC (Liskavets et al., 2025)	Llama-3.2-1B (Grattafiori et al., 2024)	1B
R2C (Choi et al., 2024)	T5-base (Raffel et al., 2020)	220M

Table 6: Summary of hyperparameters for various methods.

Hyperparameter	StreamingLLM	H2O	SnapKV	PyramidKV	Ada-SnapKV	LAQ++	SpecKV	SpecPrefill	SpecPC
Window size $n_{\text{window}}$	32	32	32	32	32	32	32	1	64
Pool	—	—	Max	Max	Max	Max	Max	Avg	Avg
Kernel size $k$	—	—	7	7	7	7	7	13	64/32
Reduction	—	—	Mean	Mean	Mean	Max	Max	Mean-Max	Max
# lookahead tokens $n_{\text{lookahead}}$	—	—	—	—	—	8	All	8	1
Compression window size $n_{\text{slash}}$	—	—	—	—	—	—	2048	—	—
# global tokens in prefill $n_{\text{vert}}$	—	—	—	—	—	—	2048	—	—
# neighbors $n_{\text{neighbor}}$ / chunk size	—	—	—	—	—	—	—	32	64/32
# skipped layers $l_{\text{skip}}$	—	—	—	—	—	—	—	—	8
Initial cache size	—	—	—	—	—	$C_{\text{max}}$	—	—	—

<sup>7</sup><https://milebench.github.io> (Apache License 2.0)

Table 5 lists the backbone models employed by each prompt compression method, while Table 6 details the hyperparameters used in our experiments. Generally, we select hyperparameters for each method based on their respective codebases. We observe that using max aggregation improved performance compared to mean aggregation for SpecKV and SpecPC. For SpecKV, setting  $n_{\text{slash}}$  and  $n_{\text{vert}}$  to 2048 resulted in minimal accuracy loss but substantially reduced latency (Section E.8.3).

For SpecKV, we always generate tokens until the draft model produces the EOS token, which yields the best performance. For latency measurements, we set  $n_{\text{lookahead}} = 64$  tokens, reflecting the average sequence length in our benchmarks. In SpecPC, prompt compression drops tokens uniformly across all layers and heads (unlike SpecKV, which prunes per head), so a larger  $C_{\text{max}}$  is needed to retain relevant information. While a larger  $n_{\text{lookahead}}$  can boost performance, in practice, generating only one token per prompt ( $n_{\text{lookahead}} = 1$ ) is usually sufficient. Strong alignment between the draft and target model attentions enables SpecPC to outperform methods like R2C and CPC. For an ablation on  $n_{\text{lookahead}}$ , see Fig. 19.

Retaining the local context for prompt compression proved essential. This observation aligns with the design of existing prompt compression methods, which typically aim to preserve entire sentences within the prompt. Consequently, we increase both the pooling kernel size ( $k$ ) and the number of neighboring tokens ( $n_{\text{neighbor}}$ ) to 64. For Llama, slightly better results are achieved by reducing both  $k$  and  $n_{\text{neighbor}}$  to 32, though the performance difference was marginal.

For all remaining methods not explicitly mentioned, we use the default configurations provided in their respective codebases.

## D.2 IMPLEMENTATION DETAILS

For our experimental results, we employ the following large language models: **Llama-3.2-1B-Instruct**<sup>8</sup>, **Llama-3.1-8B-Instruct**<sup>9</sup>, **Llama-3.1-70B-Instruct (4-bit)**<sup>10</sup>, **Qwen2.5-0.5B-Instruct**<sup>11</sup>, **Qwen2.5-14B-Instruct**<sup>12</sup>, **Qwen2.5-32B-Instruct**<sup>13</sup>, and **Qwen2.5-72B-Instruct-GPTQ-Int4**<sup>14</sup>. For MLLM evaluation on MileBench (Dingjie et al., 2024), we utilize **Qwen2.5-VL-3B-Instruct-AWQ**<sup>15</sup> and **Qwen2.5-VL-32B-Instruct-AWQ**<sup>16</sup>.

Our implementation is based on PYTORCH (Paszke et al., 2017) (BSD-3 License) and Huggingface’s TRANSFORMERS (Wolf et al., 2019) (Apache License 2.0). All experiments leverage FLASHATTENTION-2<sup>17</sup> (Dao, 2024). Latency measurements are performed using vLLM<sup>18</sup> wherever possible (i.e., where attention map outputs are not required). For implementing the sparse prefill mechanism of SpecKV, we use kernels from MINFERENCE<sup>19</sup>. All methods are evaluated via greedy decoding. Experiments are conducted on NVIDIA H100 80GB GPUs, with runtimes varying by context length; for a maximum context length of 64K tokens, experiments take up to 2 hours.

For evaluating STREAMINGLLM (Xiao et al., 2024), H2O (Zhang et al., 2023), SNAPKV (Li et al., 2024), and PYRAMIDKV (Cai et al., 2024b), we use implementations from KVCACHE-FACTORY<sup>20</sup>. In this library, the STREAMINGLLM and H2O implementations drop KV once after prefill, rather than at each decoding step, differing from their original codebases. This adjustment enables fairer comparison to SNAPKV and others. We extend KVCACHE-FACTORY to support Grouped Query Attention (Ainslie et al., 2023) by repeating keys and values for each KV head, computing attention within the window, and averaging across KV heads. This approach avoids duplicating the KV cache.

<sup>8</sup><https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct> (Llama 3.2 license)

<sup>9</sup><https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct> (Llama 3.1 license)

<sup>10</sup><https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct> (Llama 3.1 license)

<sup>11</sup><https://huggingface.co/Qwen/Qwen2.5-0.5B-Instruct> (Apache License 2.0)

<sup>12</sup><https://huggingface.co/Qwen/Qwen2.5-14B-Instruct> (Apache License 2.0)

<sup>13</sup><https://huggingface.co/Qwen/Qwen2.5-32B-Instruct> (Apache License 2.0)

<sup>14</sup><https://huggingface.co/Qwen/Qwen2.5-72B-Instruct-GPTQ-Int4> (Apache License 2.0)

<sup>15</sup><https://huggingface.co/Qwen/Qwen2.5-VL-3B-Instruct-AWQ> (Apache License 2.0)

<sup>16</sup><https://huggingface.co/Qwen/Qwen2.5-VL-32B-Instruct-AWQ> (Apache License 2.0)

<sup>17</sup><https://github.com/Dao-AILab/flash-attention> (BSD 3-Clause License)

<sup>18</sup><https://github.com/vllm-project/vllm> (Apache License 2.0)

<sup>19</sup><https://github.com/microsoft/MInference> (MIT License)

<sup>20</sup><https://github.com/Zefan-Cai/KVCache-Factory> (MIT License)

For other baselines, we use their official implementations. For Lookahead Q-Cache (LAQ++), we provide our own implementation since their code is not publicly available.

## E EXTENDED EXPERIMENTAL ANALYSIS

### E.1 ADDITIONAL RESULTS ON RULER AND LONGBENCH

In this section, we present results for RULER and LongBench using various  $C_{\max}$  values of 256, 512, and 1024 for KV dropping, and 1024, 2048, and 3072 for prompt compression. Specifically, we employ Qwen2.5-32B-Instruct, Llama-3.1-70B-Instruct (4-bit quantized with bitsandbytes<sup>21</sup>), and Qwen2.5-72B-Instruct-GPTQ-Int4 as target model.

Figs. 10 and 11 show the RULER results, and Tables 9 to 11 present the LongBench results. Overall, our methods achieve higher accuracy than the baselines in most settings, especially with small  $C_{\max}$ . Specifically, SpecKV significantly outperforms SnapKV and LAQ++ on RULER in most cases. Similarly, SpecPC consistently achieves strong results, particularly at longer sequence lengths on RULER. On LongBench, both of our methods also surpass the baselines.

### E.2 INTEGRATION AND COMPARISON WITH THE ADAKV BASELINE

Fig. 12 (RULER), Tables 12 and 13 (LongBench) present the performance of our proposed methods on Qwen2.5 (0.5B draft, 14B target) and Llama-3 (1B draft, 8B target). We include additional baselines, notably AdaKV (Feng et al., 2025), for various values of  $C_{\max}$ . AdaKV is an extension to SnapKV and allows different KV budget per attention head. We apply AdaKV to SpecKV in a similar fashion, which further boosts performance in our experiments.

### E.3 MULTI-MODAL EVALUATION

We conduct additional experiments using Qwen2.5-VL-3B-Instruct-AWQ (draft) and Qwen2.5-VL-32B-Instruct-AWQ (target) on six MileBench (Dingjie et al., 2024) datasets. We select three datasets each from the Temporal Multi-Image (EgocentricNavigation, MovingDirection, SceneTransition) and Semantic Multi-Image (SlideVQA, TQA, WebQA) categories. We focus on these datasets because, for Qwen2.5-VL, the performance gap between draft and target models is most significant; in other cases, the models perform too similarly or the draft even outperforms the target.

For KV dropping, we evaluate H2O, SnapKV, and PyramidKV from our prior experiments. We do not include AdaKV in our evaluation as it is dependent on an older Transformers (Wolf et al., 2019) version incompatible with Qwen2.5-VL. For prompt compression, we compare with FastV (Chen et al., 2024)—a method specialized for dropping image tokens inside LLMs. FastV uses a hyperparameter  $k$ : it runs all tokens up to layer  $k$ , then drops less-attended image tokens based on the attention map, processing only the top tokens thereafter. This makes FastV less efficient than SpecPC, since all tokens must be processed up to  $k$  with the full model, requiring considerable memory. Notably, FastV must compute the entire attention map at layer  $k$ , preventing the use of FlashAttention and leading to out-of-memory errors, even for moderate sequence lengths. As a result, many MileBench datasets exceed 80GB VRAM, so we limit our analysis to these six datasets.

Since the selected MileBench datasets have relatively short average context lengths, we conduct experiments using reduced  $C_{\max}$  values for both KV cache dropping (64, 96, and 128) and prompt compression (512, 768, and 1024).

Fig. 13a presents results for various KV dropping methods. Our proposed method, SpecKV, demonstrates performance comparable to existing approaches, while significantly outperforming the others on the WebQA task.

Fig. 13b compares the performance of SpecPC and FastV under two configurations ( $k = 2$  and  $k = 5$ ). Our method consistently outperforms FastV in most cases.

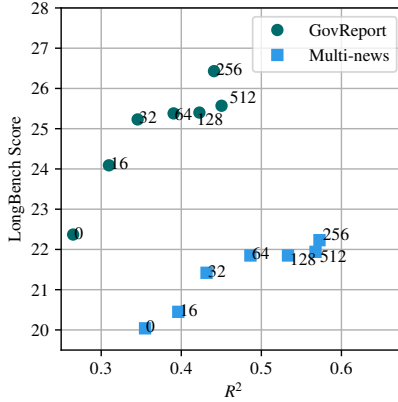


Figure 7: Impact of  $n_{\text{lookahead}}$  on SpecKV importance score accuracy ( $R^2$ ) and LongBench downstream performance ( $n_{\text{out}} = 512$ ). The plot shows a strong positive correlation: as  $n_{\text{lookahead}}$  increases (with  $n_{\text{lookahead}} = 0$  being equivalent to SnapKV), both the  $R^2$  (correlation with the true target model scores) and the downstream task score improve. Notably, SpecKV improves correlation and accuracy even with a small  $n_{\text{lookahead}} = 32$ , which is only 6.25% of the output length. Experiments use a Qwen2.5-0.5B draft model and a Qwen2.5-32B target model on two LongBench tasks.

#### E.4 SPECKV: IMPACT OF $n_{\text{LOOKAHEAD}}$ PARAMETER ON IMPORTANCE SCORE CORRELATION

We analyze the impact of the  $n_{\text{lookahead}}$  parameter on the correlation ( $R^2$ ) between SpecKV’s estimated importance scores and the ground-truth scores from the target model. For this analysis, we use 50 examples from the Multi-news and GovReport tasks from LongBench, selected for their long maximum output length (512 tokens). We employ Qwen2.5-0.5B as the draft model and Qwen2.5-32B as the target model, with  $C_{\text{max}} = 256$ .

To establish the ground-truth scores, we use the target model’s generated output (up to  $n_{\text{out}}$ ) as a perfect lookahead sequence and record the resulting importance scores. In Figs. 7, 14 and 15, we compute the  $R^2$  correlation between these ground-truth scores and the scores estimated by SpecKV using various  $n_{\text{lookahead}}$  values.

As shown in Figs. 14 and 15, the  $R^2$  correlation steadily increases with  $n_{\text{lookahead}}$ , confirming that a larger lookahead provides a more accurate importance estimation. This benefit is particularly pronounced for longer output sequences ( $n_{\text{out}}$ ). We also test  $n_{\text{lookahead}} = 0$  (equivalent to SnapKV), which yields a significantly inferior correlation.

Furthermore, we connect this score accuracy to downstream performance. Fig. 7 plots the  $R^2$  value against the final LongBench downstream score. We observe that these two metrics are highly correlated: a higher  $R^2$  (better score accuracy) generally yields better downstream performance. Consequently, increasing  $n_{\text{lookahead}}$  improves both the  $R^2$  correlation and the final task score.

#### E.5 PERFORMANCE OF CROSS-FAMILY MODELS

In this section, we evaluate the effectiveness of our methods using cross-family draft models. We set Llama-3.1-70B as the target model and employ Qwen2.5-0.5B and Qwen2.5B-1.5B as draft models. The evaluation is conducted on the RULER and LongBench benchmarks, using  $C_{\text{max}} = 256$  for SpecKV and  $C_{\text{max}} = 1024$  for SpecPC.

A key challenge in this setup is that the Qwen and Llama families use different tokenizers, necessitating a token translation step. For SpecKV, we de-tokenize the draft model’s output and then re-tokenize it with the target model’s tokenizer. For SpecPC, we aggregate attention scores from tokens to words before re-tokenizing, to avoid tokenizing partial words, and replace the draft model’s chat template with the target’s.

<sup>21</sup><https://github.com/bitsandbytes-foundation/bitsandbytes> (MIT License)

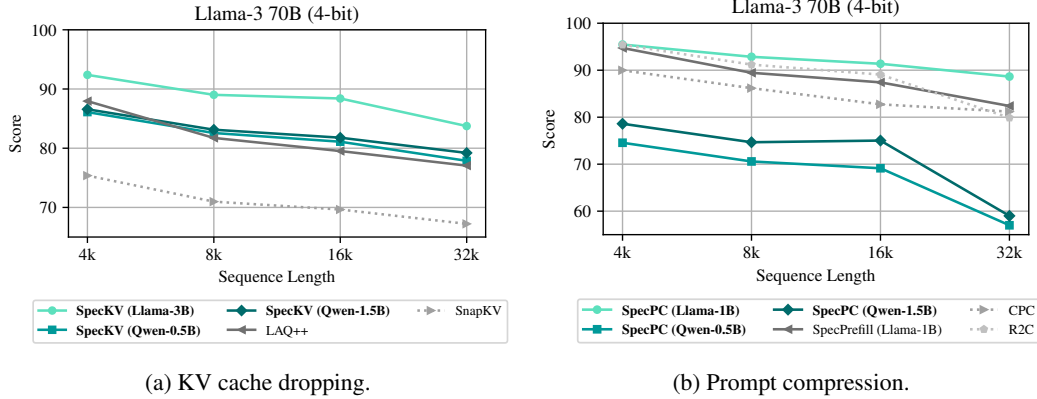


Figure 8: Cross-family model results for SpecKV and SpecPC on RULER. SpecKV demonstrates robust performance even with draft models from different families, outperforming the strongest baseline, LAQ++.

Table 7: Cross-family model results for SpecKV and SpecPC on LongBench using Llama-3.1-70B (4-bit) as target model.

	$C_{\max}$	Method	Single-doc QA	Multi-doc QA	Summary	Few-shot Learning	Code Completion	All
KV	256	SnapKV	<b>55.88</b>	45.30	22.49	62.15	55.49	47.75
		LAQ++	<u>54.90</u>	46.48	22.83	<u>64.31</u>	55.10	48.43
		<b>SpecKV (Llama-3B)</b>	51.80	<b>47.23</b>	<b>25.53</b>	64.02	<b>58.75</b>	<b>48.80</b>
		<b>SpecKV (Qwen-0.5B)</b>	53.18	45.70	24.40	63.69	57.36	48.26
		<b>SpecKV (Qwen-1.5B)</b>	51.20	<u>47.07</u>	<u>24.98</u>	<b>65.81</b>	<u>57.53</u>	<u>48.73</u>
PC	1024	CPC	45.14	39.41	24.86	61.40	37.58	41.97
		R2C	48.93	42.01	25.38	58.91	40.19	43.29
		SpecPrefill (Llama-1B)	<u>54.62</u>	<b>46.43</b>	25.63	64.80	44.92	<u>48.37</u>
		<b>SpecPC (Llama-1B)</b>	<b>56.84</b>	44.48	<b>25.91</b>	<b>67.37</b>	47.15	<b>48.44</b>
		<b>SpecPC (Qwen-0.5B)</b>	36.73	36.42	24.68	64.51	<b>50.75</b>	42.04
		<b>SpecPC (Qwen-1.5B)</b>	51.58	39.53	<u>25.87</u>	<u>66.43</u>	<u>50.25</u>	46.48

The results, shown in Fig. 8 (RULER) and Table 7 (LongBench), indicate that both SpecKV and SpecPC achieve good results even with cross-family drafts. Notably, SpecKV outperforms the strongest baseline, LAQ++. This aligns with the intuition that SpecKV, which relies only on the draft model’s output, should generalize well to cross-model scenarios. SpecPC experiences a performance drop because it relies on similar attention patterns between the draft and target models, a condition that may not hold in cross-family setups.

## E.6 EXTENDED RESULTS FOR SPECKV-PC

This section presents extended results for the cascaded compression strategy, SpecKV-PC. By restricting the large target model to process only a small fraction of the original prompt, this approach achieves substantially better latency and memory efficiency than SpecKV alone.

We evaluate the impact of varying prompt compression ratios on accuracy using Qwen2.5-1.5B/32B and Llama-3-3B/70B as draft/target pairs, with a final KV cache size of  $C_{\max} = 256$ . Our combined method, denoted as SpecKV-PC-X (where the prompt is compressed to X tokens), is benchmarked against standard SpecKV, SpecPC, LAQ++, and SnapKV.

As illustrated in Fig. 9 (RULER) and Table 8 (LongBench), SpecKV-PC proves to be both efficient and highly accurate. Notably, we find that moderate pre-compression (e.g., SpecKV-PC-2048) yields accuracy superior to standard SpecKV alone, particularly at the longer sequence lengths (RULER). This suggests that the initial SpecPC stage acts as an effective pre-filter, discarding irrelevant information to help the target model focus on the most important tokens.

Conversely, extreme prompt compression (e.g., SpecPC-256) results in significant performance degradation. This indicates that for aggressive compression targets, coarse prompt-level reduction is insufficient.

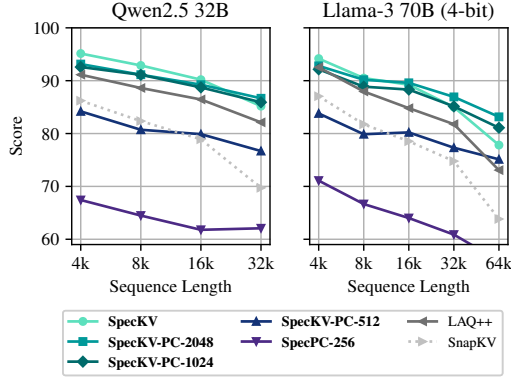


Figure 9: Performance of the combined SpecKV and SpecPC methods on RULER (final  $C_{\max} = 256$ ). Our cascaded approach (SpecKV-PC-X) pre-compresses the prompt to X tokens, achieving higher accuracy than standard SpecKV (which sees the full prompt) at longer sequences. This suggests the initial pre-filtering allows SpecKV to make a more effective final selection.

Table 8: LongBench performance of cascaded compression with SpecKV and SpecPC using Qwen2.5 and Llama-3. SpecKV-PC-2048 achieves superior performance.

		Qwen2.5 32B						Llama-3 70B (4-bit)					
Group	Method	SingleQA	MultiQA	Summ.	Few-shot	Code	All	SingleQA	MultiQA	Summ.	Few-shot	Code	All
Dense	Target	56.01	43.99	25.90	64.06	44.74	47.78	55.02	47.06	28.61	70.47	48.19	49.99
KV	SnapKV	52.54	40.21	19.89	61.18	40.12	42.98	55.88	45.30	22.49	62.15	55.49	47.75
	LAQ++	55.15	44.14	22.24	63.25	41.19	45.79	54.90	46.48	22.83	64.31	55.10	48.43
	SpecKV	53.48	43.77	24.02	<b>63.79</b>	44.80	46.06	51.80	47.23	25.53	64.02	<b>58.75</b>	48.80
	SpecKV-PC-2048	52.60	<b>44.52</b>	24.11	63.38	<b>48.45</b>	<b>46.48</b>	<b>61.42</b>	47.15	<b>26.51</b>	<b>66.94</b>	58.19	<b>51.60</b>
	SpecKV-PC-1024	<b>55.37</b>	42.21	<b>24.14</b>	63.09	39.72	45.28	<u>58.73</u>	<b>47.33</b>	<u>25.69</u>	64.28	55.19	<u>49.89</u>
	SpecKV-PC-512	45.77	36.57	22.36	56.57	39.57	40.21	51.73	46.11	23.01	63.64	50.00	46.68
PC	SpecPC-256	34.75	27.05	19.15	45.0	35.07	32.0	27.97	25.93	19.09	51.25	48.12	33.5

## E.7 SPECKV: EXTENDED LATENCY AND MEMORY ANALYSIS

We analyze the latency and memory usage of baselines and our algorithms, breaking down latency into three stages: draft generation, target prefill, and target decoding. For LAQ++, draft generation occurs after target prefill to share the same KV cache.

All experiments are run on an H200 GPU (141 GB VRAM) using the Qwen2.5 model (3B draft, 32B target). We evaluate across a range of input sizes,  $n_{\text{in}} \in \{4k, 8k, 16k, 32k, 64k\}$ , and output sizes,  $n_{\text{out}} \in \{64, 128, 256, 512\}$ . A global  $C_{\max} = 256$  is used for all experiments. For algorithm-specific settings, we set  $n_{\text{lookahead}} = n_{\text{out}}$  for SpecKV and use the default  $n_{\text{lookahead}} = 8$  for LAQ++. For the SpecKV-PC variant, we employ SpecKV-PC-2048, which pre-compresses the prompt to 2048 tokens.

Our results demonstrate that draft lookahead introduces negligible latency (Fig. 16) and memory overhead (Fig. 17) relative to the total cost of target model inference. Notably, cascaded compression with SpecKV and SpecPC (SpecKV-PC) significantly reduces prefill time and peak memory. At 64k context, SpecKV-PC is about 40% faster and 25 GB more memory efficient than LAQ++.

## E.8 ABLATION STUDIES

In this section, we conduct a series of ablation experiments to further analyze the effectiveness of our two proposed methods: SpecKV and SpecPC. For consistency, we fix  $C_{\max}$  to 256 for KV dropping



and 1024 for prompt compression across all experiments. We utilize Qwen2.5 (Instruct), employing the 0.5B model as the draft and the 14B model as the target. We sample 100 random examples per task from the LongBench and RULER benchmarks.

#### E.8.1 SPECKV: ANALYSIS OF ENHANCED DRAFT MODELS

Fig. 18 illustrates the impact of using draft models of different sizes and versions on the RULER score. As anticipated, both newer (Yang et al., 2025a) and larger draft models lead to improved performance of SpecKV.

#### E.8.2 SPECKV AND SPECPC: IMPACT OF $n_{\text{LOOKAHEAD}}$

Fig. 19 illustrates how varying the number of generated draft tokens,  $n_{\text{lookahead}}$ , affects the performance of SpecKV and SpecPC. Overall, increasing  $n_{\text{lookahead}}$  generally results in higher final accuracy for SpecKV, whereas it yields only marginal improvements for SpecPC. We attribute this to the larger  $C_{\text{max}}$  budget of SpecPC, which allows it to capture all important tokens without needing to generate long drafts.

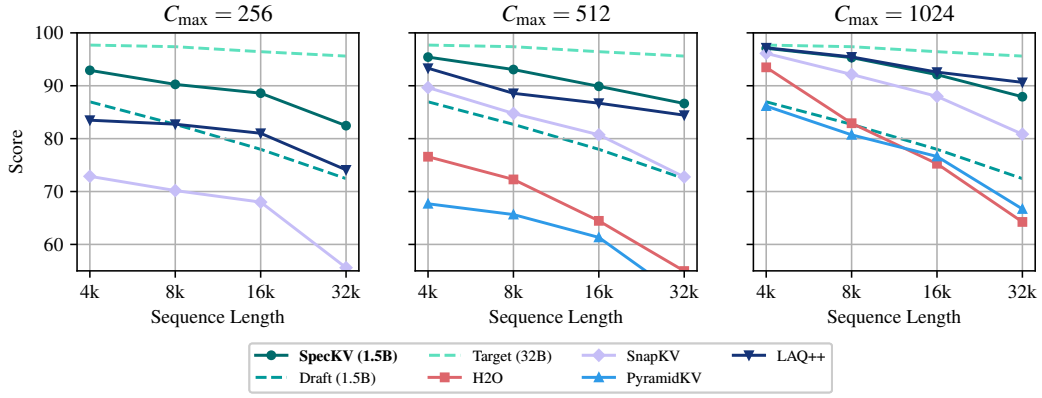
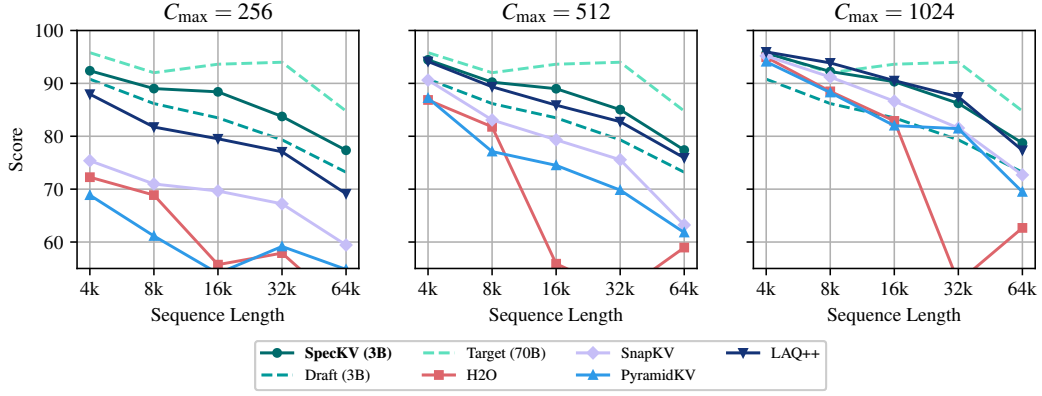
#### E.8.3 SPECKV: ACCURACY ANALYSIS OF SPARSE PREFILL

In this section, we experimentally evaluate how the sparsity of SpecKV’s prefill procedure affects downstream task performance (Fig. 20). Specifically, we set  $n_{\text{vert}}$  equal to  $n_{\text{slash}}$  and compare several values for these parameters. As anticipated, reducing sparsity (i.e., using a higher  $n_{\text{vert}}$ ) generally results in higher accuracy; however, accuracy improvements plateau at  $n_{\text{vert}} = 2048$ , which we therefore adopt for our main experiments.

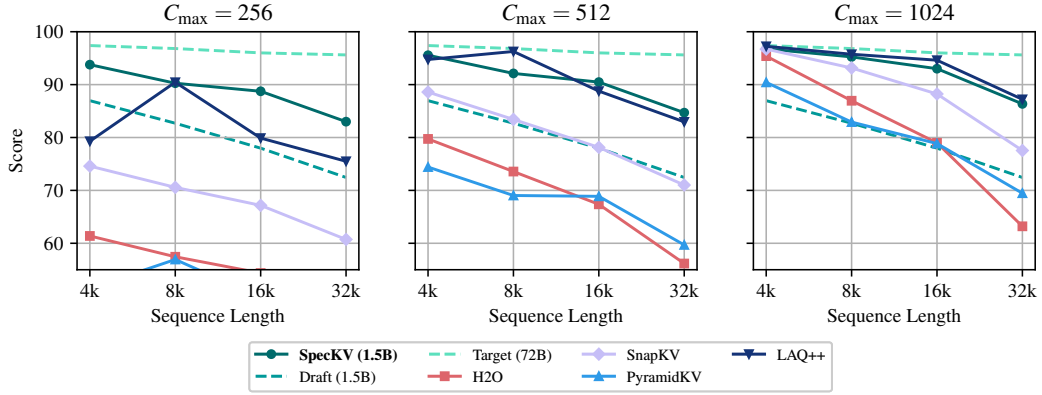
Interestingly, for certain LongBench categories, increased sparsity (i.e., lower  $n_{\text{vert}}$ ) can actually lead to improved performance. This counterintuitive result suggests that, for some tasks, sparser prefill may serve as a form of regularization, preventing overfitting to irrelevant context.

### LLM USAGE DISCLOSURE

We utilized Google’s Gemini 2.5 Pro (Google DeepMind, 2025) and OpenAI’s GPT (Achiam et al., 2023) to assist with improving the grammar, clarity, and readability of this manuscript. The authors reviewed and edited all LLM-generated suggestions to ensure the final text accurately reflects our scientific contributions and claims. The authors retain full responsibility for the content of this paper.

(a) Qwen2.5: Draft model (1.5B), target model (32B)<sup>22</sup>

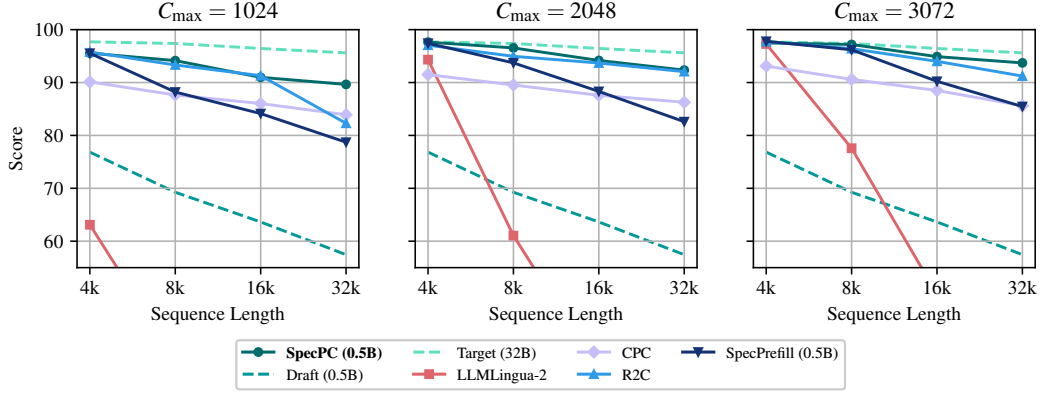
(b) Llama-3: Draft model (3.2-3B), target model (3.1-70B, 4-bit)



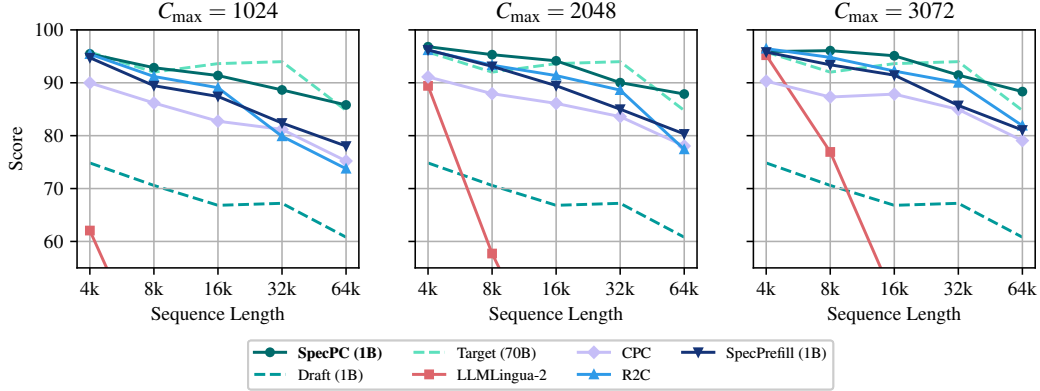
(c) Qwen2.5: Draft model (1.5B), target model (72B, 4-bit)

Figure 10: Extended RULER results for KV cache dropping. Our proposed SpecKV method consistently outperforms SnapKV and LAQ++ across the majority of evaluated settings, often by a substantial margin.

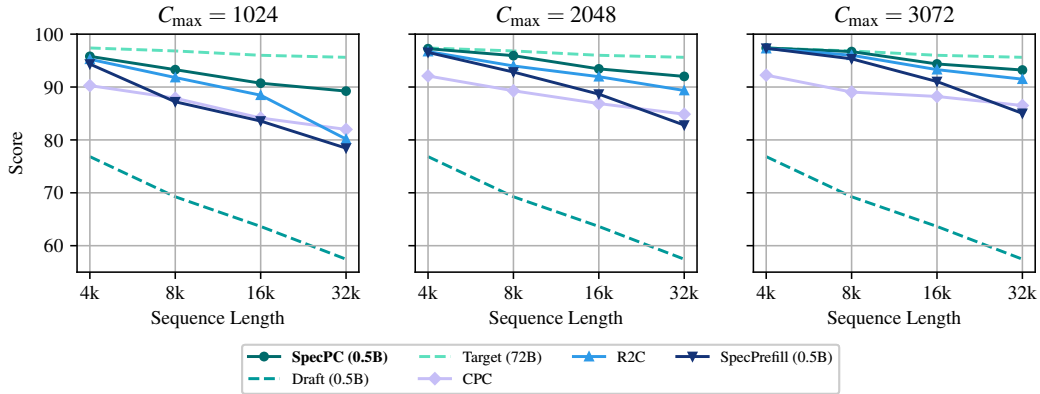
<sup>22</sup>H2O and PyramidKV are not plotted for Qwen2.5 at  $C_{\max} = 256$  as their performance falls outside the visible range.



(a) Qwen2.5: Draft model (0.5B), target model (32B)



(b) Llama-3: Draft model (3.2-1B), target model (3.1-70B, 4-bit)



(c) Qwen2.5: Draft model (0.5B), target model (72B, 4-bit)

Figure 11: Extended RULER results on prompt compression. Our proposed SpecPC method consistently outperforms CPC, R2C, and SpecPrefill, maintaining strong performance even on long sequences.

Table 9: LongBench results for Qwen2.5, featuring 0.5B (SpecPC) and 1.5B (SpecKV) draft models and a 32B target model.

	$C_{\max}$	Method	Single-doc QA	Multi-doc QA	Summary	Few-shot Learning	Code Completion	All
Dense	–	Draft (0.5B)	19.07	26.58	20.90	53.51	32.48	30.37
		Draft (1.5B)	36.16	35.01	22.79	63.92	36.62	39.06
		Target (32B)	56.01	43.99	25.90	64.06	44.74	47.78
KV	256	H2O	46.63	30.81	19.88	56.03	39.27	39.32
		SnapKV	52.54	40.21	19.89	61.18	40.12	42.98
		PyramidKV	50.92	37.26	18.90	63.24	40.20	43.19
		LAQ++	<b>55.15</b>	<b>44.14</b>	<u>22.24</u>	<u>63.25</u>	<u>41.19</u>	<u>45.79</u>
		<b>SpecKV (1.5B)</b>	<u>53.48</u>	<u>43.77</u>	<b>24.02</b>	<b>63.79</b>	<b>44.80</b>	<b>46.06</b>
	512	H2O	48.62	36.72	21.31	59.10	40.99	42.27
		SnapKV	55.24	42.21	21.47	<u>63.36</u>	39.69	44.73
		PyramidKV	54.79	41.87	20.28	62.00	40.08	<u>46.55</u>
		LAQ++	<b>55.77</b>	<b>45.30</b>	<u>23.96</u>	62.37	<u>41.86</u>	46.44
		<b>SpecKV (1.5B)</b>	52.78	<u>43.97</u>	<b>24.80</b>	<b>64.72</b>	<b>46.77</b>	<b>46.60</b>
	1024	H2O	51.86	41.30	23.02	60.18	42.93	44.84
		SnapKV	55.32	44.04	23.08	<b>66.15</b>	<b>44.42</b>	46.76
		PyramidKV	<u>55.91</u>	42.67	22.28	<u>64.76</u>	42.52	<b>48.31</b>
		LAQ++	<b>56.33</b>	<b>45.18</b>	<u>25.17</u>	62.13	<u>43.06</u>	46.61
		<b>SpecKV (1.5B)</b>	55.72	43.95	<b>25.20</b>	63.31	42.38	46.38
	1024	CPC	45.60	<u>40.62</u>	23.09	60.08	32.31	40.91
		R2C	<u>50.49</u>	40.37	<u>23.26</u>	53.45	34.11	39.88
		SpecPrefill (0.5B)	45.94	39.32	23.16	<u>62.04</u>	<b>43.17</b>	<u>42.70</u>
		<b>SpecPC (0.5B)</b>	<b>51.23</b>	<b>41.40</b>	<b>23.37</b>	<b>62.26</b>	<u>38.23</u>	<b>43.66</b>
PC	2048	CPC	51.01	42.31	23.74	60.92	35.83	43.26
		R2C	50.32	<b>42.66</b>	24.08	59.11	40.54	44.19
		SpecPrefill (0.5B)	<u>51.60</u>	40.72	23.71	<b>64.20</b>	<u>45.29</u>	<u>45.09</u>
		<b>SpecPC (0.5B)</b>	<b>55.40</b>	<u>42.60</u>	<b>24.12</b>	<u>61.46</u>	<b>48.05</b>	<b>46.20</b>
	3072	CPC	<u>56.80</u>	42.05	<b>24.77</b>	<u>62.79</u>	37.98	45.37
		R2C	51.67	<u>42.88</u>	24.09	62.45	27.01	42.66
		SpecPrefill (0.5B)	53.85	<b>42.92</b>	24.48	<b>64.03</b>	45.76	<u>46.24</u>
		<b>SpecPC (0.5B)</b>	<b>56.89</b>	42.42	<u>24.71</u>	62.66	<b>47.49</b>	<b>46.79</b>

Table 10: LongBench results for Llama-3, featuring 3.2-1B (SpecPC) and 3.2-3B (SpecKV) draft models and a 3.1-70B (4-bit) target model.

	$C_{\max}$	Method	Single-doc QA	Multi-doc QA	Summary	Few-shot Learning	Code Completion	All
Dense	–	Draft (1B)	28.37	28.62	26.12	59.10	33.59	35.27
		Draft (3B)	45.53	40.52	27.46	64.82	46.73	44.89
		Target (70B)	55.02	47.06	28.61	70.47	48.19	49.99
KV	256	H2O	54.07	41.30	22.55	49.10	54.14	43.52
		SnapKV	<b>55.88</b>	45.30	22.49	62.15	55.49	47.75
		PyramidKV	<u>55.41</u>	45.59	22.50	59.06	49.90	46.25
		LAQ++	54.90	<u>46.48</u>	<u>22.83</u>	<b>64.31</b>	55.10	<u>48.43</u>
		<b>SpecKV (3B)</b>	51.80	<b>47.23</b>	<b>25.53</b>	<u>64.02</u>	<b>58.75</b>	<b>48.80</b>
	512	H2O	55.01	43.86	23.85	58.48	52.00	46.26
		SnapKV	<b>56.08</b>	46.94	24.40	63.34	<u>52.14</u>	48.32
		PyramidKV	54.73	47.07	24.20	<u>63.62</u>	47.05	47.35
		LAQ++	<u>55.06</u>	<u>47.45</u>	<u>24.58</u>	63.39	51.75	<u>48.36</u>
		<b>SpecKV (3B)</b>	<u>53.31</u>	<b>47.47</b>	<b>27.33</b>	<b>67.49</b>	<b>54.19</b>	<b>49.66</b>
	1024	H2O	54.19	45.77	26.03	63.62	<u>49.58</u>	47.71
		SnapKV	54.94	<b>47.40</b>	25.97	<b>67.56</b>	48.13	<b>48.85</b>
		PyramidKV	<b>56.32</b>	46.82	26.20	62.58	48.28	48.02
		LAQ++	<u>55.19</u>	<u>47.17</u>	<u>26.43</u>	<u>67.41</u>	47.16	48.61
		<b>SpecKV (3B)</b>	53.32	46.02	<b>27.41</b>	66.40	<b>50.66</b>	<u>48.63</u>
	1024	CPC	45.14	39.41	24.86	61.40	37.58	41.97
		R2C	48.93	42.01	25.38	58.91	40.19	43.29
		SpecPrefill (1B)	<u>54.62</u>	<b>46.43</b>	<u>25.63</u>	<u>64.80</u>	<u>44.92</u>	<u>48.37</u>
		<b>SpecPC (1B)</b>	<b>56.84</b>	<u>44.48</u>	<b>25.91</b>	<b>67.37</b>	<b>47.15</b>	<b>48.44</b>
	2048	CPC	55.97	46.00	26.72	64.78	41.31	47.36
		R2C	53.62	<b>46.70</b>	26.27	62.41	<b>47.90</b>	47.34
		SpecPrefill (1B)	<u>58.39</u>	45.37	27.13	<u>66.60</u>	45.45	<u>48.81</u>
		<b>SpecPC (1B)</b>	<b>59.39</b>	<u>46.25</u>	<b>27.60</b>	<b>68.42</b>	<u>46.70</u>	<b>49.88</b>
	3072	CPC	56.64	46.69	27.29	<u>64.92</u>	<u>44.75</u>	<u>48.30</u>
		R2C	56.11	45.15	<u>27.51</u>	61.76	<b>47.17</b>	<u>47.57</u>
		SpecPrefill (1B)	<u>57.75</u>	46.15	27.15	64.70	42.50	48.02
		<b>SpecPC (1B)</b>	<b>58.47</b>	<b>48.07</b>	<b>27.72</b>	<b>65.43</b>	41.28	<b>48.69</b>

Table 11: LongBench results for Qwen2.5, featuring 0.5B (SpecPC) and 1.5B (SpecKV) draft models and a 72B (4-bit) target model.

	$C_{\max}$	Method	Single-doc QA	Multi-doc QA	Summary	Few-shot Learning	Code Completion	All
Dense	–	Draft (0.5B)	19.07	26.58	20.90	53.51	32.48	30.37
		Draft (1.5B)	36.16	35.01	22.79	63.92	36.62	39.06
		Target (72B)	58.70	45.89	26.24	64.83	51.08	49.22
KV	256	H2O	53.68	35.05	21.10	50.84	48.87	41.41
		SnapKV	55.85	40.12	21.65	55.82	50.41	44.37
		PyramidKV	55.13	38.89	20.14	53.02	49.59	42.91
		LAQ++	<b>57.26</b>	<b>43.54</b>	<u>22.95</u>	<u>62.53</u>	<b>55.79</b>	<u>46.98</u>
		<b>SpecKV (1.5B)</b>	53.94	<u>42.71</u>	<b>25.07</b>	<b>66.58</b>	46.93	<b>47.06</b>
	512	H2O	55.77	39.52	22.61	58.01	50.69	44.94
		SnapKV	<b>59.08</b>	<b>44.13</b>	23.04	61.33	<u>51.77</u>	<u>47.59</u>
		PyramidKV	56.98	42.52	22.03	55.43	50.29	45.10
		LAQ++	58.28	34.94	<u>24.38</u>	<u>63.74</u>	<b>54.79</b>	<b>49.08</b>
		<b>SpecKV (1.5B)</b>	55.74	<u>43.39</u>	<b>25.85</b>	<b>64.56</b>	47.80	47.44
	1024	H2O	55.37	43.50	24.11	62.09	50.68	46.90
		SnapKV	<b>59.56</b>	44.06	24.41	<u>62.85</u>	<u>52.87</u>	<u>48.46</u>
		PyramidKV	58.34	43.68	23.18	59.66	51.45	46.96
		LAQ++	<u>58.66</u>	35.37	<b>26.35</b>	60.90	<b>53.56</b>	<b>49.59</b>
		<b>SpecKV (1.5B)</b>	56.04	<b>44.27</b>	<u>25.71</u>	<b>63.46</b>	49.90	47.73
	1024	CPC	46.37	38.36	24.19	52.54	28.30	38.64
		R2C	<b>55.78</b>	40.10	<u>24.40</u>	47.62	33.17	40.72
		SpecPrefill (0.5B)	<u>49.04</u>	<u>40.32</u>	<b>24.41</b>	<b>63.53</b>	<b>50.12</b>	<u>45.16</u>
		<b>SpecPC (0.5B)</b>	48.52	<b>45.34</b>	24.14	<u>61.04</u>	<u>48.30</u>	<b>45.26</b>
PC	2048	CPC	55.14	43.53	24.71	49.46	33.21	41.78
		R2C	54.61	<u>45.70</u>	24.99	53.65	42.43	44.41
		SpecPrefill (0.5B)	53.39	43.46	<b>25.53</b>	<b>65.89</b>	<u>50.18</u>	<u>47.51</u>
		<b>SpecPC (0.5B)</b>	<b>58.36</b>	<b>46.38</b>	<u>25.32</u>	<u>64.43</u>	<b>51.11</b>	<b>48.98</b>
	3072	CPC	55.78	44.41	25.47	42.71	39.16	41.68
		R2C	<u>59.22</u>	44.70	25.69	55.22	44.04	45.90
		SpecPrefill (0.5B)	55.75	<b>45.71</b>	<b>25.93</b>	<b>64.22</b>	<b>51.63</b>	<u>48.44</u>
		<b>SpecPC (0.5B)</b>	<b>59.84</b>	<u>44.95</u>	<u>25.91</u>	<u>64.18</u>	<u>46.88</u>	<b>48.46</b>

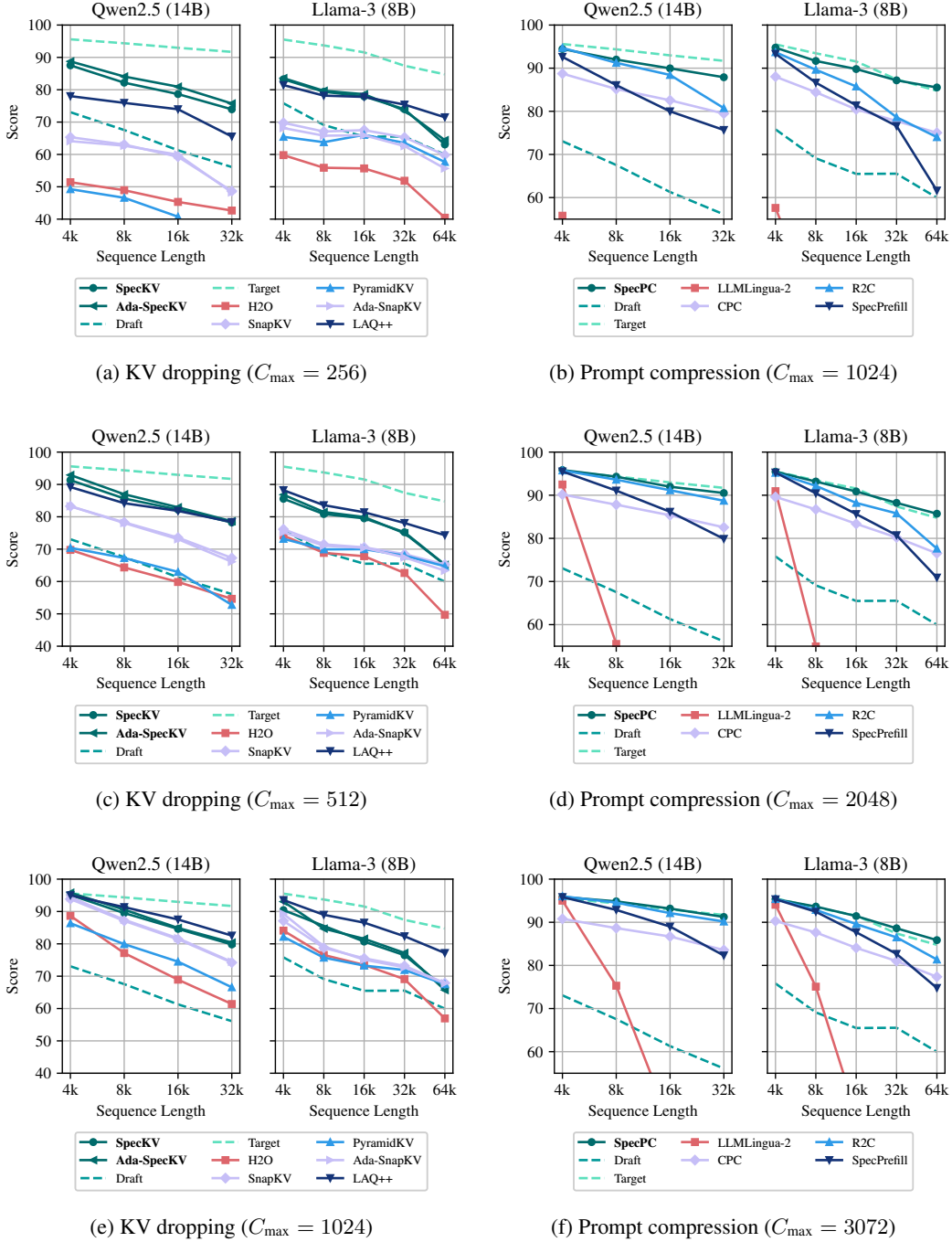


Figure 12: Performance of the proposed SpecKV and SpecPC on RULER compared to various baselines including AdaKV for Qwen2.5-0.5B-Instruct and Llama-3.2-1B-Instruct as draft models, and Qwen2.5-14B-Instruct and Llama-3.1-8B-Instruct as target models. AdaKV can further boost the performance of SpecKV. Notably, SpecPC performs nearly on par with the target model. Low-performing methods are omitted.

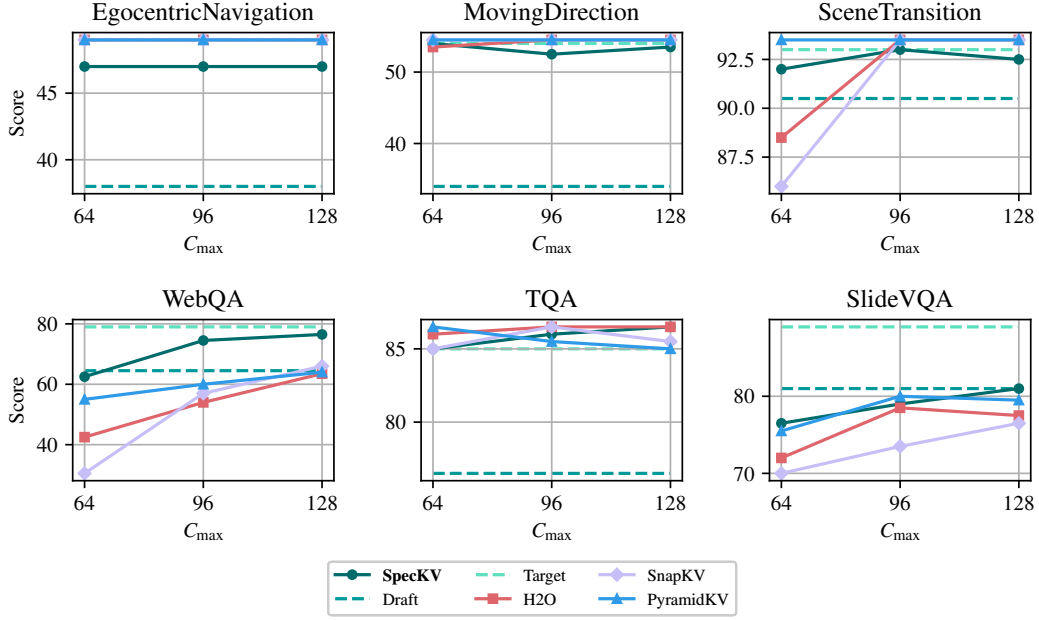


Table 12: Results for LongBench performance with Qwen2.5-0.5B (draft) and Qwen2.5-14B (target).

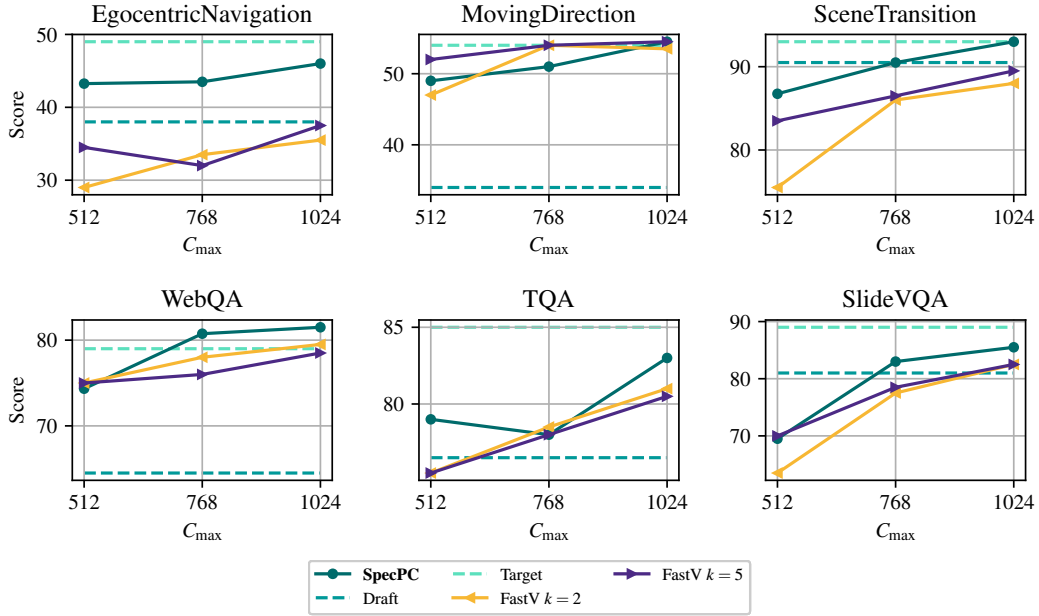
	$C_{\max}$	Method	Single-doc QA	Multi-doc QA	Summary	Few-shot Learning	Code Completion	All
Dense	–	Draft	21.04	24.4	21.23	54.07	33.39	30.64
		Target	53.19	42.83	24.99	65.34	51.75	47.33
KV	256	StreamingLLM	38.66	24.13	19.10	49.75	35.23	33.24
		H2O	46.98	29.66	19.82	50.88	47.86	38.41
		SnapKV	49.07	33.19	19.49	54.49	47.34	40.25
		PyramidKV	47.07	31.84	18.32	54.81	43.26	38.76
		Ada-SnapKV	50.92	34.57	19.63	55.07	49.58	41.41
		LAQ++	<b>52.45</b>	<b>39.99</b>	22.12	<b>61.46</b>	51.32	45.05
		<b>SpecKV</b>	51.07	38.76	<u>23.20</u>	59.68	49.58	44.09
		<b>Ada-SpecKV</b>	<u>51.92</u>	<u>39.38</u>	<b>24.90</b>	<u>61.05</u>	<b>53.43</b>	<b>45.61</b>
	512	StreamingLLM	38.97	27.04	21.14	53.78	36.52	35.41
		H2O	47.56	33.16	21.02	53.45	49.78	40.37
		SnapKV	50.74	38.33	20.97	57.81	49.94	43.10
		PyramidKV	50.19	36.37	20.25	58.17	47.88	42.19
		Ada-SnapKV	51.69	38.61	21.48	59.84	51.85	44.18
		LAQ++	<b>52.31</b>	<u>42.06</u>	23.75	61.62	51.64	45.89
		<b>SpecKV</b>	<u>51.98</u>	<u>41.51</u>	<u>24.11</u>	<u>62.60</u>	<u>52.33</u>	46.09
		<b>Ada-SpecKV</b>	51.00	<b>42.18</b>	<b>25.74</b>	<b>63.74</b>	<b>55.28</b>	<b>48.46</b>
	1024	StreamingLLM	40.08	32.01	22.12	55.42	38.36	37.54
		H2O	49.73	37.27	22.38	55.94	51.26	42.75
		SnapKV	51.56	40.98	22.65	63.90	51.48	45.73
		PyramidKV	51.42	40.93	21.76	59.96	49.92	44.43
		Ada-SnapKV	51.81	40.96	22.83	<u>64.14</u>	51.97	45.94
		LAQ++	<u>52.27</u>	<u>42.59</u>	<u>25.04</u>	61.56	51.68	46.27
		<b>SpecKV</b>	<b>52.92</b>	41.85	24.30	63.46	<u>52.50</u>	46.61
		<b>Ada-SpecKV</b>	<u>52.27</u>	<b>43.22</b>	<b>26.39</b>	<b>64.83</b>	<b>54.40</b>	<b>47.78</b>
PC	1024	LLMLingua-2	27.10	23.74	22.57	35.85	37.62	28.79
		CPC	42.76	36.05	22.58	48.38	35.57	37.17
		R2C	<u>46.43</u>	<u>37.42</u>	22.64	47.21	29.48	37.15
		SpecPrefill	46.21	37.21	<u>22.82</u>	<b>62.18</b>	48.34	43.00
		<b>SpecPC</b>	<b>47.70</b>	<b>38.30</b>	<b>23.30</b>	<u>59.74</u>	<b>52.52</b>	<b>43.73</b>
	2048	LLMLingua-2	35.10	32.22	23.51	39.82	44.80	34.40
		CPC	48.74	39.70	23.39	55.77	42.07	41.92
		R2C	<u>50.68</u>	<u>40.65</u>	23.47	56.72	45.62	43.27
		SpecPrefill	49.75	38.58	<b>23.70</b>	<b>65.05</b>	50.40	45.15
		<b>SpecPC</b>	<b>53.23</b>	<b>41.43</b>	<u>23.68</u>	<u>63.26</u>	<b>54.92</b>	<b>46.76</b>
	3072	LLMLingua-2	43.50	35.67	24.10	45.66	47.32	38.67
		CPC	51.59	41.08	23.84	58.14	44.82	43.83
		R2C	51.71	<u>41.39</u>	23.89	62.09	48.58	45.31
		SpecPrefill	<u>51.81</u>	41.27	<u>24.26</u>	<u>64.28</u>	50.66	<u>46.15</u>
		<b>SpecPC</b>	<b>53.25</b>	<b>41.48</b>	<b>24.49</b>	<b>64.51</b>	<b>54.41</b>	<b>47.14</b>

Table 13: Results for LongBench performance with Llama-3.2-1B (draft) and Llama-3.1-8B (target).

	$C_{\max}$	Method	Single-doc QA	Multi-doc QA	Summary	Few-shot Learning	Code Completion	All
Dense	–	Draft	28.08	27.27	25.65	60.16	31.11	34.69
		Target	45.85	43.79	28.68	66.65	50.46	46.84
KV	256	StreamingLLM	38.69	27.12	21.64	50.75	34.74	34.58
		H2O	43.54	36.81	22.62	55.64	47.81	40.82
		SnapKV	43.79	37.31	21.96	56.29	47.34	40.91
		PyramidKV	43.62	37.79	21.75	55.15	46.32	40.54
		Ada-SnapKV	44.04	37.86	22.34	59.95	50.39	42.38
		LAQ++	<b>45.81</b>	<b>41.72</b>	22.93	<b>64.18</b>	47.21	<b>44.17</b>
		<b>SpecKV</b>	43.23	39.73	<u>24.43</u>	<u>60.90</u>	<u>51.09</u>	<u>43.36</u>
		<b>Ada-SpecKV</b>	42.40	<u>40.73</u>	<b>25.74</b>	57.94	<b>52.51</b>	43.25
	512	StreamingLLM	38.80	29.15	24.16	52.59	37.25	36.33
		H2O	44.82	39.36	23.95	58.70	49.30	42.79
		SnapKV	45.00	41.31	23.61	61.36	49.92	43.83
		PyramidKV	<u>45.26</u>	41.22	23.36	61.15	48.06	43.51
		Ada-SnapKV	45.09	41.17	24.19	61.81	<u>51.74</u>	44.30
		LAQ++	<b>46.17</b>	<b>43.44</b>	24.17	<b>63.13</b>	48.74	<b>44.87</b>
		<b>SpecKV</b>	43.48	41.86	<u>26.17</u>	<u>62.34</u>	51.36	<u>44.59</u>
		<b>Ada-SpecKV</b>	43.80	<u>42.56</u>	<b>26.59</b>	58.72	<b>54.41</b>	44.56
	1024	StreamingLLM	38.45	32.54	25.03	58.09	38.60	38.54
		H2O	45.45	42.50	25.42	59.17	50.10	44.13
		SnapKV	45.46	43.15	25.42	62.06	<u>52.37</u>	45.22
		PyramidKV	<u>46.10</u>	42.78	25.17	63.43	<u>49.55</u>	45.11
		Ada-SnapKV	46.06	43.34	25.48	<u>63.47</u>	50.50	45.43
		LAQ++	<b>46.21</b>	<b>44.14</b>	25.71	<b>64.10</b>	50.00	<b>45.75</b>
		<b>SpecKV</b>	43.73	43.39	<u>26.95</u>	63.14	51.28	45.30
		<b>Ada-SpecKV</b>	44.83	<u>43.72</u>	<b>27.50</b>	61.04	<b>54.23</b>	<u>45.70</u>
PC	1024	LLMLingua-2	29.61	24.83	23.43	24.18	40.66	27.68
		CPC	35.67	36.61	25.26	34.01	43.58	34.42
		R2C	38.41	39.07	25.28	43.26	43.99	37.58
		SpecPrefill	<u>41.51</u>	39.38	<u>25.82</u>	<u>63.51</u>	44.68	42.86
		<b>SpecPC</b>	<b>44.83</b>	<b>39.94</b>	<b>25.85</b>	<b>63.70</b>	<b>44.82</b>	<b>43.76</b>
	2048	LLMLingua-2	34.00	32.51	24.90	24.76	<b>47.27</b>	31.64
		CPC	40.02	39.41	26.83	39.02	46.66	37.80
		R2C	<u>44.53</u>	38.97	26.63	54.62	46.67	41.97
		SpecPrefill	42.35	<b>42.11</b>	<u>27.23</u>	<u>63.56</u>	44.46	43.91
		<b>SpecPC</b>	<b>44.92</b>	<u>40.71</u>	<b>27.30</b>	<b>64.77</b>	<u>46.89</u>	<b>44.78</b>
	3072	LLMLingua-2	39.13	35.44	25.98	29.73	<b>49.92</b>	35.05
		CPC	41.73	39.52	27.27	42.12	<u>49.13</u>	39.30
		R2C	<u>44.77</u>	40.97	27.35	60.85	48.08	44.14
		SpecPrefill	43.86	<b>42.45</b>	<u>27.82</u>	<u>63.17</u>	45.25	<u>44.46</u>
		<b>SpecPC</b>	<b>47.12</b>	<u>41.95</u>	<b>28.02</b>	<b>65.61</b>	45.52	<b>45.65</b>



(a) KV dropping



(b) Prompt compression

Figure 13: MileBench multi-modal results using Qwen2.5-VL-3B-Instruct-AWQ (draft) and Qwen2.5-VL-32B-Instruct-AWQ (target). SpecKV demonstrates competitive performance across most tasks and achieves a substantial improvement on WebQA. SpecPC consistently outperforms both FastV configurations on the majority of datasets.



Figure 14: Correlation of importance scores from SpecKV against ground-truth scores from the dense target model for various output lengths ( $n_{\text{out}}$ ) and lookaheads ( $n_{\text{lookahead}}$ ). Increasing  $n_{\text{lookahead}}$  improves correlation, especially for longer output lengths ( $n_{\text{out}}$ ). SnapKV ( $n_{\text{lookahead}} = 0$ ) consistently shows the lowest correlation. Experiments use a Qwen2.5-0.5B draft model and a Qwen2.5-32B target model on Multi-news from LongBench.



Figure 15: Correlation of importance scores from SpecKV against ground-truth scores from the dense target model for various output lengths ( $n_{\text{out}}$ ) and lookaheads ( $n_{\text{lookahead}}$ ). Increasing  $n_{\text{lookahead}}$  improves correlation, especially for longer output lengths ( $n_{\text{out}}$ ). SnapKV ( $n_{\text{lookahead}} = 0$ ) shows the lowest correlation. Experiments use a Qwen2.5-0.5B draft model and a Qwen2.5-32B target model on GovReport from LongBench.

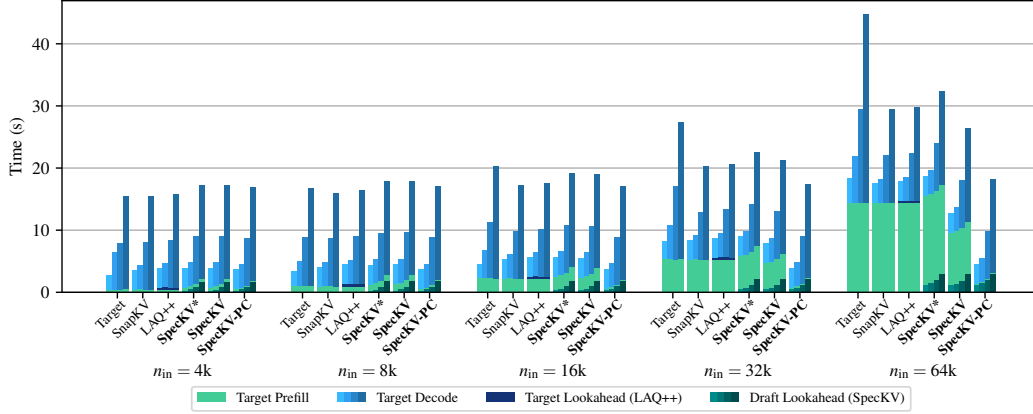


Figure 16: End-to-end latency comparison across five input sequence lengths ( $n_{in}$ ) and four output sequence lengths ( $n_{out} \in \{64, 128, 256, 512\}$ ). All experiments use the Qwen2.5 model (3B draft, 32B target) with  $C_{max} = 256$ . For SpecKV algorithms, we use  $n_{lookahead} = n_{out}$ . Each bar segment breaks down the latency by processing stage, and the four bars for each method correspond to the four  $n_{out}$  values. SpecKV\* denotes SpecKV without sparse prefill, while SpecKV-PC combines SpecKV with SpecPC. Notably, when  $n_{in} \geq 16k$ , the efficiency gains from sparse prefill (SpecKV) and prompt compression (SpecKV-PC) effectively offset the lookahead overhead for  $n_{lookahead} \leq 64$  and  $\leq 512$ , respectively. Overall, SpecKV-PC is the most efficient method; by precompressing the prompt, it significantly cuts target prefill time, resulting in a speedup of about 40% to 75% (depending on output length) over LAQ++ at a 64k input context.

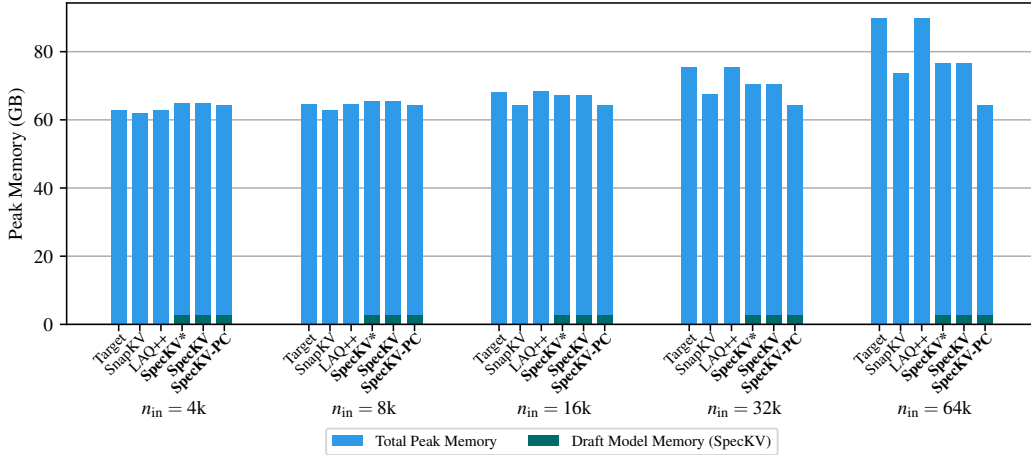


Figure 17: Peak memory usage across five input sequence lengths ( $n_{in}$ ) using the Qwen2.5 model (3B draft, 32B target) with  $C_{max} = 256$  and  $n_{lookahead} = n_{out}$ . SpecKV\* denotes SpecKV without sparse prefill, while SpecKV-PC combines SpecKV with SpecPC. LAQ++ offers no peak memory savings over the target model, as it must store the full cache. SpecKV is slightly more memory-hungry than SnapKV because it also stores the draft model weights (indicated in green), though this constitutes a small fraction of the total memory. SpecKV-PC is the most memory-efficient method; it first compresses the prompt (to 2048 tokens here) before the target model’s prefill phase, significantly reducing peak usage. At a 64k context length, SpecKV-PC saves approximately 9 GB over SnapKV and 25 GB over LAQ++.

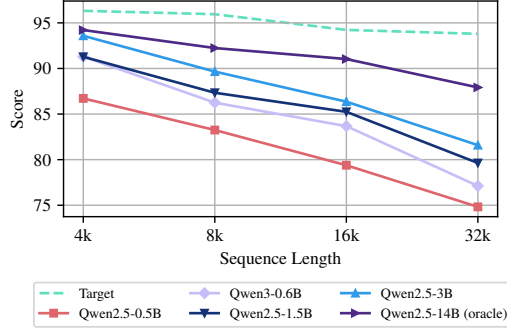


Figure 18: Effect of draft model quality on RULER score. The target model is Qwen2.5-14B (Instruct). Consistent with expectations, employing more capable draft models boosts performance. For reference, we also evaluate an oracle setting where the draft model is identical to the target (Qwen2.5-14B), representing an empirical upper bound for SpecKV.

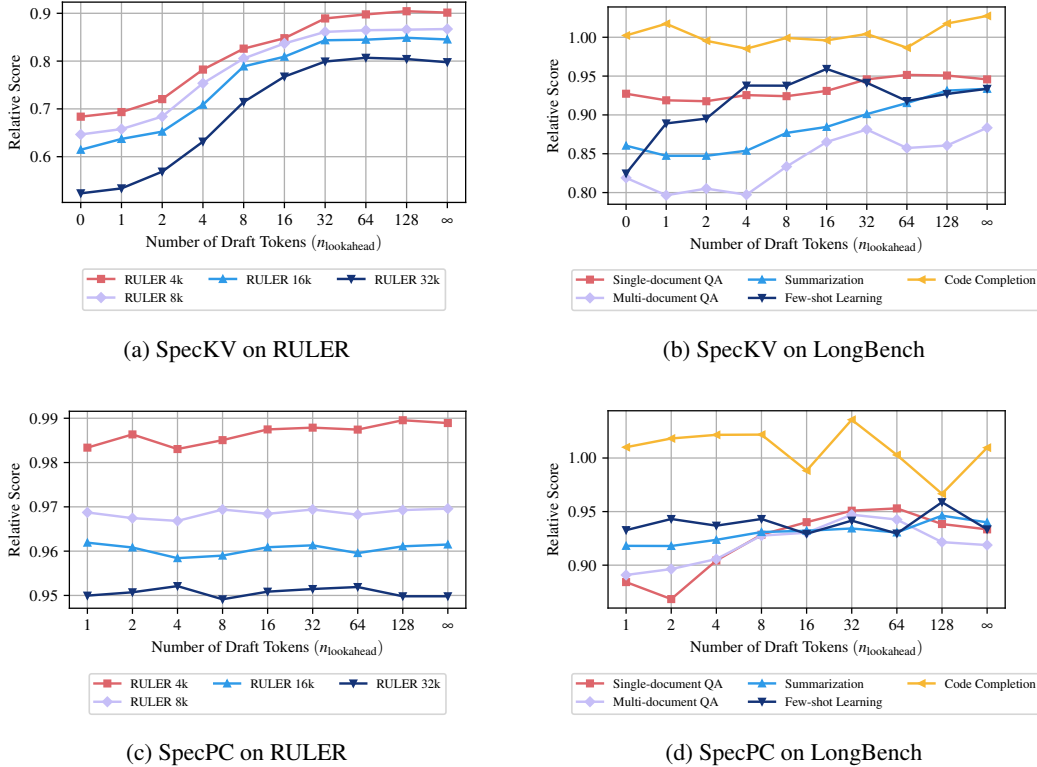


Figure 19: Impact of the number of generated draft tokens,  $n_{\text{lookahead}}$ , on the relative performance of SpecKV and SpecPC, using Qwen2.5-0.5B-Instruct as the draft model and Qwen2.5-14B-Instruct as the target model. The relative score is calculated as the score of each SpecKV configuration divided by the score of the full dense target model. Increasing  $n_{\text{lookahead}}$  substantially boosts SpecKV’s score, whereas SpecPC shows only minor improvement with higher  $n_{\text{lookahead}}$ . SnapKV ( $n_{\text{lookahead}} = 0$ ) has the lowest performance in most cases.  $\infty$  denotes lookahead to the EOS token.

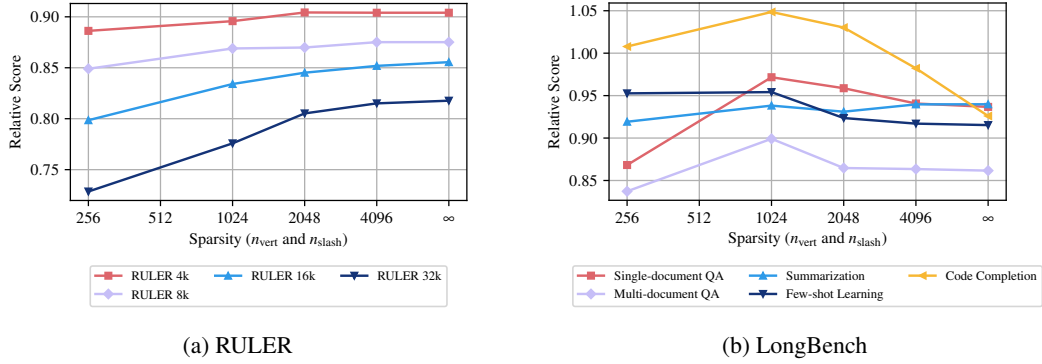


Figure 20: Impact of different sparsity levels in SpecKV’s sparse prefill on relative performance, using Qwen2.5-0.5B-Instruct as the draft model and Qwen2.5-14B-Instruct as the target model. We vary  $n_{\text{vert}}$  (set equal to  $n_{\text{slash}}$ ) and observe the impact on accuracy. Accuracy improves as sparsity decreases (higher  $n_{\text{vert}}$ ) up to 2048, beyond which gains saturate, hence our choice of 2048 for main results.  $\infty$  corresponds to fully dense prefill. Notably, for some LongBench tasks, higher sparsity actually benefits accuracy. The relative score is calculated as the score of each SpecKV configuration divided by the score of the full dense target model.