D. Additional Related Work

KV cache sharing Recent work has explored various strategies for sharing KV caches across transformer layers. Layer-Condensed KV Cache (LCKV) (Wu & Tu, 2024) computes KVs only for the top layer and pairs them with queries from all layers, while optionally retaining standard attention for a few top and bottom layers to mitigate performance degradation. Similarly, You Only Cache Once (YOCO) (Sun et al., 2024) computes KVs exclusively for the top layer but pairs them with queries from only the top half of layers, employing efficient attention in the bottom layers to maintain a constant cache size. In contrast, Cross-Layer Attention (CLA) (Brandon et al., 2024) divides layers into groups, pairing queries from all layers in each group with KVs from that group's bottom layer. MiniCache (Liu et al., 2024a) introduces a novel method that merges layer-wise KV caches while enabling recovery during compute-in-place operations, optimizing KV cache size. These methods illustrate various trade-offs between computation, memory usage, and model performance when sharing KV caches across transformer layers.

Long-Context Benchmarks The landscape of long-context model benchmarks has evolved to encompass a wide range of tasks, with particular emphasis on retrieval and comprehension capabilities. Benchmarks for understanding have made significant strides, with ∞-Bench (Zhang et al., 2024) pushing the boundaries by presenting challenges that involve more than 100,000 tokens. LongBench (Bai et al., 2024) has introduced bilingual evaluations, addressing tasks such as long-document question answering, summarization, and code completion. Complementing these efforts, ZeroSCROLLS (Shaham et al., 2023) and L-Eval (An et al., 2023) have broadened the scope to include a diverse array of practical natural language tasks, including query-driven summarization.

In parallel, retrieval benchmarks have largely relied on synthetic datasets, offering researchers precise control over variables such as the length of input tokens. This approach minimizes the impact of disparate parametric knowledge resulting from varied training methodologies. A significant body of recent work has concentrated on the development of synthetic tasks specifically for retrieval evaluation (Kamradt, 2023; Mohtashami & Jaggi, 2023; Li et al., 2023; Liu et al., 2024c; Hsieh et al., 2024). In addition, researchers have explored the potential of extended contexts in facilitating various forms of reasoning (Tay et al., 2021).

This dual focus on synthetic retrieval tasks and comprehensive understanding benchmarks reflects the field's commitment to rigorously assessing the capabilities of long-context models across diverse linguistic challenges. **Prompting Compression** In the field of prompt compression, various designs effectively combine semantic information to compress natural language. Wingate et al. (2022) utilize soft prompts to encode more information with fewer tokens. Chevalier et al. (2023) present AutoCompressor, which uses soft prompts to compress the input sequence and extend the original length of the base model. Both Zhou et al. (2023) and Wang et al. (2023) recurrently apply LLMs to summarize input texts, maintaining long short-term memory for specific purposes such as story writing and dialogue generation. The LLMLingua series (Jiang et al., 2023b; 2024; Fei et al., 2024) explores the potential of compressing LLM prompts in long-context, reasoning, and RAG scenarios. Fei et al. (2024) use pre-trained language models to chunk the long context and summarize semantic information, compressing the original context.

E. Statistics of Models

Table 18 provides configuration parameters for LLMs that we evaluated in our experiments.

Model Name	LLaMA-3-8B-Instruct	Mistral-7B-Instruct-v0.2 & 0.3	Qwen2-7B-Instruct
L (Number of layers)	32	32	28
N (Number of attention heads)	32	32	28
D (Dimension of each head)	128	128	128

Table 18: Models Configuration Parameters