A APPENDIX

A.1 Qualitative Comparison of Text Generation

This section qualitatively and quantitatively compares text generation during summarization tasks using various KV cache reduction methods on the MPT-7B (MosaicML, 2023) pre-trained model, focusing on Keyformer. The qualitative assessment is based on a random sample from the CNN/DailyMail (See et al., 2017) validation dataset, and ROUGE scores (Lin, 2004) are employed for quantitative comparisons. Keyformer performs better than other techniques with similar KV cache reduction.

Input: *Douglas Costa will spark a transfer scramble this summer with Shakhtar Donetsk ready to sell their prized-asset. Chelsea manager Jose Mourinho is a known admirer of the Brazil international having tried to land the midfielder in the previous transfer windows. Shakhtar chiefs are now open to selling Costa this summer and talks with third parties over his departure are underway. Brazil international Douglas Costa is set to depart Shakhtar Donetsk for £25million in the summer . Midfielder Costa (left) could spark a bidding war from Chelsea, Real Madrid and Barcelona . The 24-year-old Costa wants to remain playing for a Champions League side and is open to a move to England . And the Ukrainian side hope to spark a bidding war for the midfielder in attempt to maximise his transfer fee, with a starting price of £25million. Costa is open to joining a Barclays Premier League club, but wants to be playing Champions League football. However, a host of Europe's top sides will also be in the running to land Costa this summer, Real Madrid and Barcelona among them. Arsenal have had Costa watched extensively, while Manchester United and Liverpool have both been linked with a move. Chelsea manager Jose Mourinho is known to be an admirer of the midfield player Costa . Costa has dual Brazilian and Portuguese nationality, removing any potential work permit problems . Chelsea are set to lead the charge for the 24-year-old having tracked the Brazil international for two years. Despite playing for Brazil, Costa has dual Portuguese nationality so should not face work permit issues.*

Reference Summary: *Shakhtar Donetsk could sell Douglas Costa for at least £25million. Jose Mourinho is a known admirer of the Brazil international midfielder. Real Madrid, Barcelona and Arsenal are also interested in the 24-year-old. READ: Douglas Costa dreaming of summer move to Chelsea .*

MPT-7B Full Attention: *Douglas Costa will spark a transfer scramble this summer with Shakhtar Donetsk ready to sell their prized-asset. Chelsea manager Jose Mourinho is a known admirer of the Brazil international having tried to land the midfielder in the previous transfer windows. Shakhtar chiefs are now open to selling Costa this summer and talks with third parties over his departure are underway. Brazil international Douglas Costa is set to depart Shakhtar Donetsk for £25million in the summer . Midfielder Costa (left) could spark a bidding war from Chelsea, Real Madrid and Barcelona .*
ROUGE-1: 49.635
ROUGE-2: 25.1852
ROUGE-L: 30.6569

MPT-7B Window Attention (50% KV Cache): *Chelsea manager Jose Mourinho is known to be an admirer of the midfield player Costa . Chelsea have had Costa watched extensively, while Manchester United and Liverpool have both been linked with a move. Chelsea have had Costa watched extensively, while Manchester United and Liverpool have both been linked with a move. Arsenal have had Costa watched extensively, while Manchester United and Liverpool have both been linked with a move. Chelsea have had Costa watched extensively, while Manchester United and Liverpool have both been linked with a move. Chelsea are set to lead the charge for the 24-year-old having tracked the Brazil international for two years.*
ROUGE-1: 31.5789
ROUGE-2: 12.0
ROUGE-L: 19.7368

MPT-7B H_2O (50% KV Cache): *Douglas Costa will spark a transfer scramble this summer with Shakhtar Donetsk ready to sell their prized asset. Chelsea manager Jose Mourinho is a known admirer of the Brazil international having tried to land the midfielder in the previous transfer windows. Shakhtar chiefs are now open to selling Costa this summer and talks over his departure are underway. Brazil international Douglas Costa will leave Shakhtar Donetsk for £25million in the summer . The 24-year-old midfielder wants to remain playing for a Champions League side and is open to a move to England .*
ROUGE-1: 44.6043
ROUGE-2: 24.8175
ROUGE-L: 33.0935

MPT-7B Keyformer (50% KV Cache:) *Douglas Costa will spark a transfer scramble this summer with Shakhtar Donetsk ready to sell their prized-asset. Chelsea manager Jose Mourinho is a known admirer of the Brazil international having tried to land the midfielder in the previous transfer windows. Shakhtar chiefs are now ready to sell Costa this summer and talks over a move to the Premier League are already underway. Brazil international Douglas Costa is set to leave Shakhtar Donetsk this summer with Chelsea, Barcelona and Real Madrid all interested in his signature. The 24-year-old has been linked with a move to the Premier League.*
ROUGE-1: 46.1538
ROUGE-2: 28.3688
ROUGE-L: 33.5664

A.2 Extended Related Work

KV Cache Compression and Quantization: Extensive research has explored model compression and quantization to accommodate larger models within limited memory. For autoregressive LLMs, as sequence lengths increase, the size of KV cache exceeds the model parameters' size. Implementing compression and quantization techniques on KV cache leads to faster inference. FastGen (Ge et al., 2024) introduces adaptive KV cache compression by preserving special tokens. ALISA (Zhao et al., 2024) utilizes sparse window attention and quantization for KV cache. GEAR (Kang et al., 2024) proposes a KV cache compression framework using quantization and low-rank matrices to approximate quantization error. QAQ (Dong et al., 2024) introduces quality-adaptive quantization for KV cache. DejaVuLib (Strati et al., 2024) implements efficient prompt-token disaggregation to reduce pipeline delays for distributed LLM serving.

Popular Embeddings in Other Large Models: The concept of identifying popular or key embeddings (and potentially dropping them) has being investigated for recommendation and other large models (Adnan et al., 2021; 2024; Wang et al., 2023). This concept is used to enable an intelligent embedding placement across heterogeneous memory hierarchies. This helps enhance training efficiency. Similarly, the presence of key tokens tokens within autoregressive LLMs aids in reducing the KV cache size and its allocation within limited GPU memory.

A.3 Sparsity within Large Language Models

We analyzed the MPT-7B model to explore the sparsity of LLMs. Our approach involved examining threshold-based sparsity by varying the percentage threshold of the maximum attention score. We tested thresholds from 0% to 5%, where 0% represents tokens with no attention score. The resulting Figure 11 illustrates increased sparsity as the percentage threshold rises.

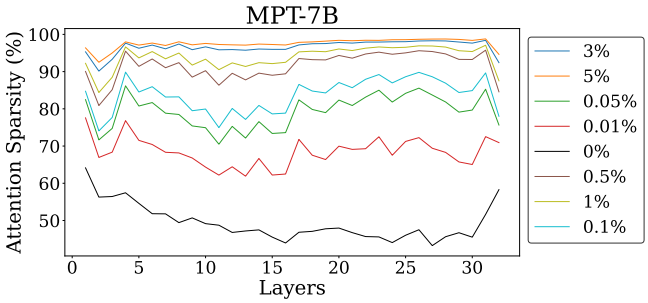

Figure 11. Increase in sparsity with varying threshold percentage.

A.4 Recent Window versus Key Token Window Ratio

We conducted a sensitivity study to examine the impact of varying the ratio of recent tokens w on the size of the KV cache dedicated to recent tokens and key tokens. We maintained a fixed KV cache size of 70%. This resulted in changes in the number of key tokens $(k - w)$. Figure 12 displays the trend in model accuracy for all three respective models. The results indicate that the models perform better when the recent tokens ratio w falls within the range of 20% to 30%. This observation aligns with our hypothesis that both recent and key tokens tokens tend to be of the highest importance for text generation tasks.

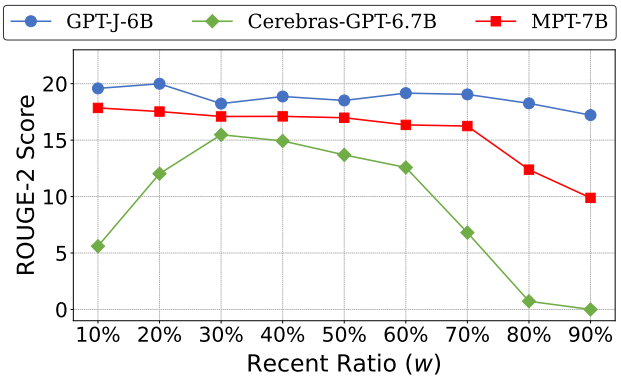

Figure 12. Varying the recent ratio (w) in Keyformer at 70% KV cache and its impact on model accuracy for summarization task with CNN/DailyMail dataset.

A.5 ROUGE-1 and ROUGE-L Scores

The MLPerf benchmarks (Reddi et al., 2020) establish rigorous standards for summarization tasks, mandating that all ROUGE scores, encompassing *ROUGE-1, ROUGE-2, and ROUGE-L*, should range between 99% and 99.9% of the original scores. Figure 13 shows the trends of ROUGE-1 and ROUGE-L scores in summarization tasks using the CNN/DailyMail validation dataset for three models.

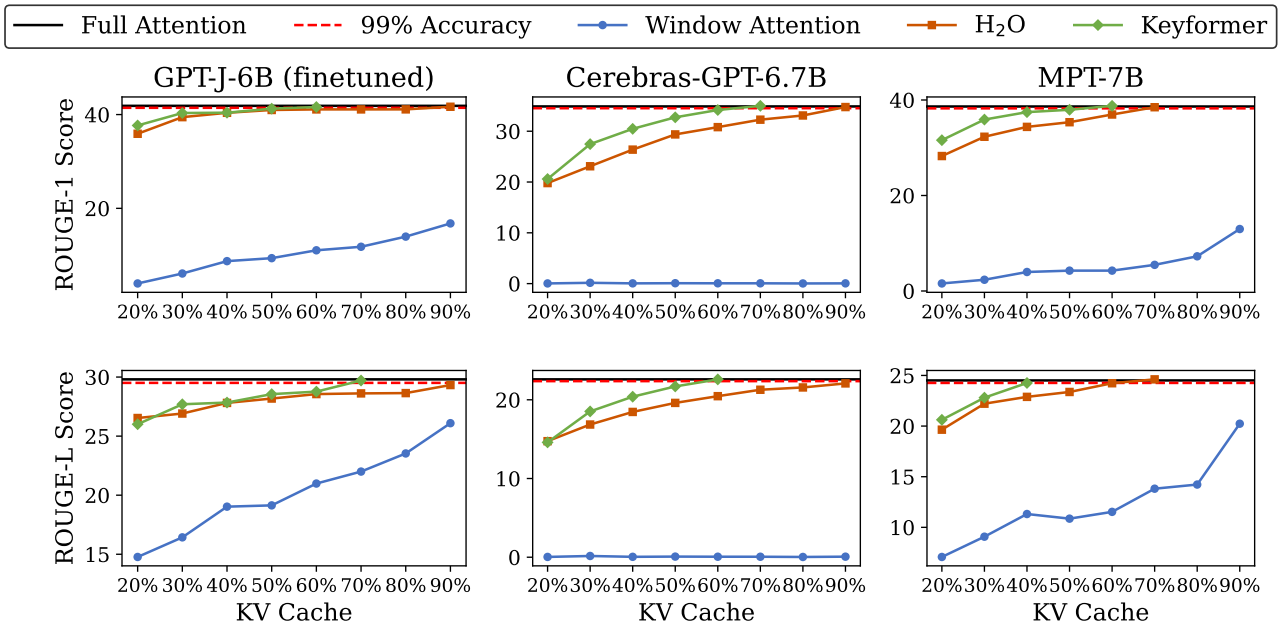

Figure 13. ROUGE-1 and ROUGE-L scores comparison of Full Attention, Window Attention, H_2O and Keyformer with varying KV cache size. The solid black line shows Full Attention without discarding any token and full KV cache. The red dotted line shows 99% accuracy mark as per MLPerf (Reddi et al., 2020).

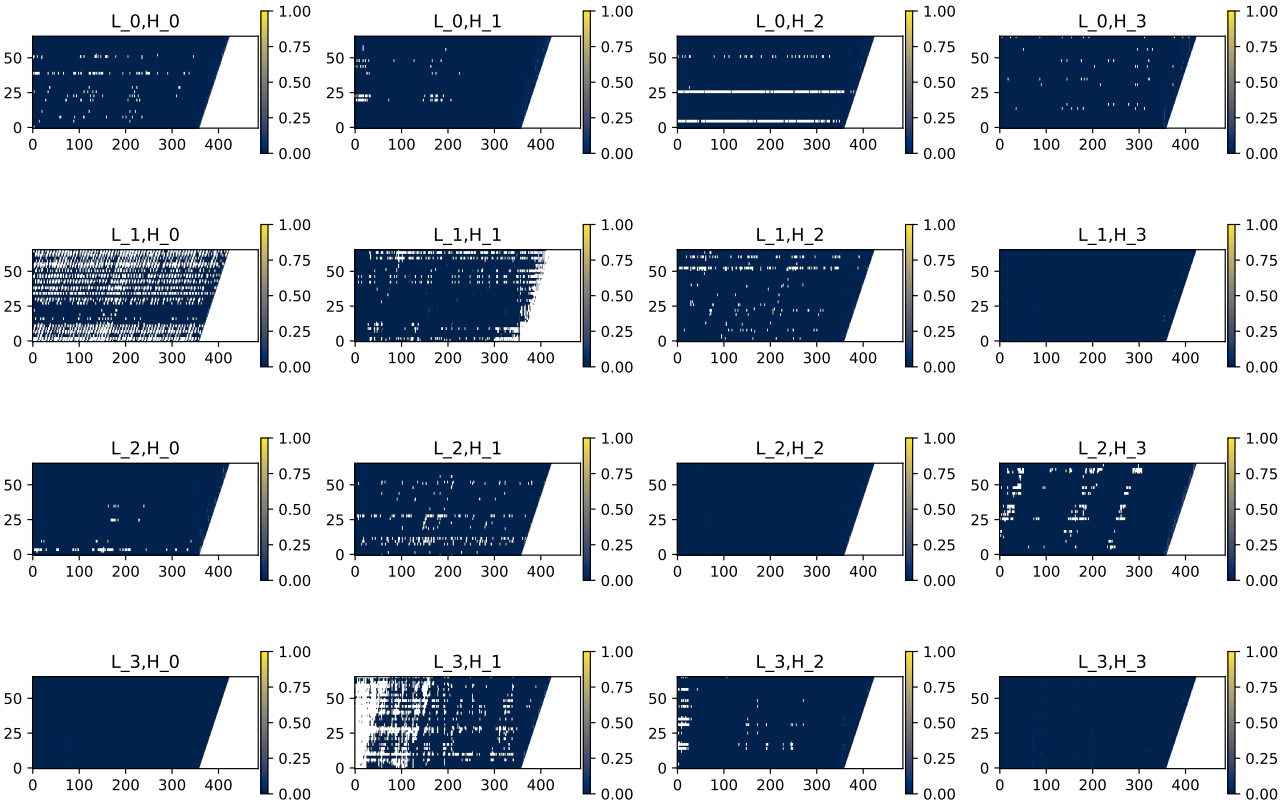

Figure 14. Attention heat map for GPT-J (Wang & Komatsuzaki, 2021) model with first 4 layers and heads (denoted as L_<layer>,H_<head>). The x-axis comprises context + text generation, while the y-axis only contains text generation. Empty (white) dots show zero attention score and present inherent sparsity.

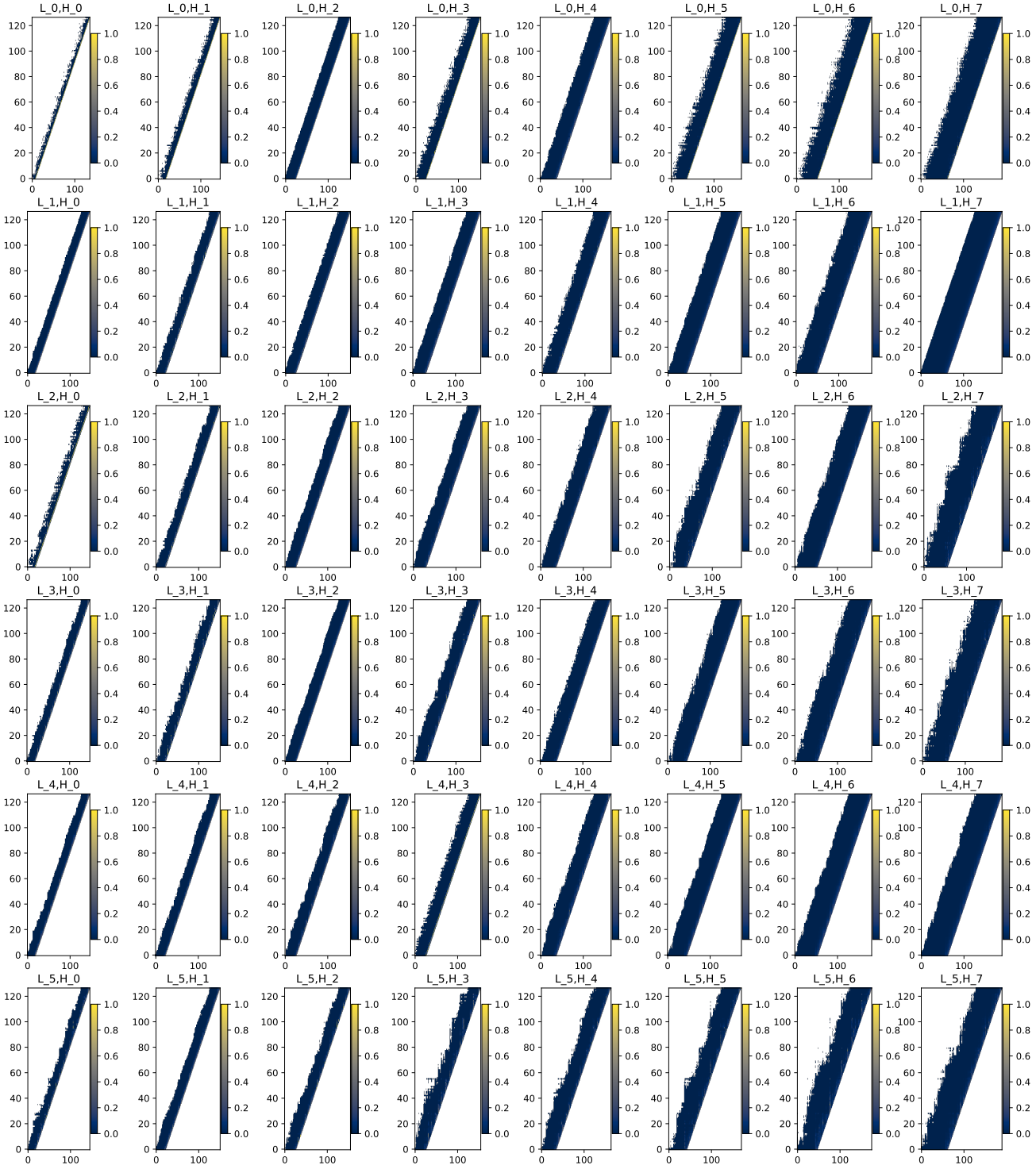

Figure 15. Attention heat map for MPT-7B (MosaicML, 2023) model with first 6 layers and 8 heads (denoted as L_<layer>,H_<head>). The x-axis comprises context + text generation, while the y-axis only contains text generation. Empty (white) dots show zero attention score and present inherent sparsity. The effect of ALiBi (Press et al., 2021) can be seen as we move from Head 0 to Head 7.

A.6 Attention Heat Maps of Large Language Models

We examined attention heat maps from different models to explore their inherent sparsity. Figure 14 presents the attention heat map for a fine-tuned GPT-J (Wang & Komatsuzaki, 2021) model using a sample from the CNN/DailyMail dataset. The input context comprises 360 tokens, and the model produces a summary of 65 tokens. The heat map reveals that inherent sparsity is dispersed across layers and heads without a distinct pattern. This lack of pattern poses a challenge for exploiting sparsity.

Figure 14 and Figure 15 depict attention heat maps across different layers and heads for the fine-tuned GPT-J and pre-trained MPT models. The variations observed in the attention heat maps can be attributed mainly to differences in the models' positional encoding schemes.

A.7 Analyzing Attention Sinks

StreamingLLM (Xiao et al., 2023) introduced the concept of "Attention Sink," which prioritizes the initial tokens deemed most important. This approach, which retains the first four tokens alongside the recent window, aims to reduce perplexity effectively. However, as indicated by the attention heat maps presented earlier, *we did not observe a similar trend in our models*. As explored in ALiBi (Press et al., 2021), the MPT model experimented with training on shorter sequence lengths and testing on longer sequences within the training domain. A constant bias was introduced into the attention mechanism to facilitate the successful generation of longer sequences. This bias gradually diminishes in value as it extends to older generated tokens, with recent tokens receiving a higher bias and tokens further back in the sequence receiving a smaller bias. This bias distribution is observable in the MPT attention heat maps. This characteristic may contribute to the sub-optimal performance of StreamingLLM in summarization tasks.

A.8 Sensitivity of the Temperature Parameter

We conducted an ablation study to examine the impact of the temperature parameter τ on Keyformer's performance. This study focused on the summarization task using the MPT-7B model and the CNN/DailyMail dataset. We varied the values of τ for the Gumbel softmax, maintaining a consistent τ throughout both the prompt processing and text generation phases. This ensured a uniform level of randomness in the Gumbel softmax distribution, independent of discarded tokens. Moreover, we adjusted the randomness intuitively to match the number of discarded tokens, aiming to compensate for them during text generation iterations. Figure 16 shows how sweeping τ levels helped identify key tokens and enhance the quality of the model compared to maintaining a fixed τ value.

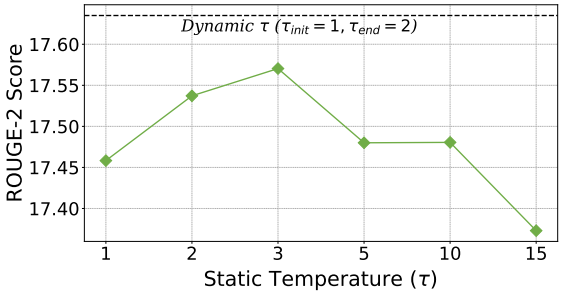

Figure 16. Varying the temperature parameter (τ) for MPT-7B model for Keyformer and its impact on model accuracy for summarization task with CNN/DailyMail dataset. Sweeping value of τ from 1 to 2 across the token generation phase works better in comparison to fixed temperature across the prompt processing and token generation phase.

A.9 Artifact Evaluation

This artifact provides the source code for Keyformer.

Access to Artifact: The artifact is available on GitHub at the following link: https://github.com/d-matrix-ai/keyformer-llm.

Datasets: For extensive evaluation across both summarization and conversation tasks, we utilized multiple datasets:

CNN/DailyMail [2]: dataset comprises over 300,000 unique news articles written by journalists from CNN and the Daily Mail. It facilitates both extractive and abstractive summarization.

GovReport [3]: dataset contains approximately 19.5k U.S. government reports featuring expert-written abstractive summaries. It includes notably longer documents (9.4k words) and summaries (553 words) compared to other existing datasets.

SODA [4]: dataset is a high-quality collection of dialogues encompassing diverse social interactions.

The validation set of each dataset is utilized for evaluation.

System Requirements: For evaluation, the following hardware is employed:

- 180 GB of disk space
- 90 GB of DRAM
- NVIDIA Tesla A100 (80 GB) GPU

[2] https://huggingface.co/datasets/cnn_dailymail

[3] https://huggingface.co/datasets/ccdv/govreport-summarization?row=0

[4] https://huggingface.co/datasets/allenai/soda