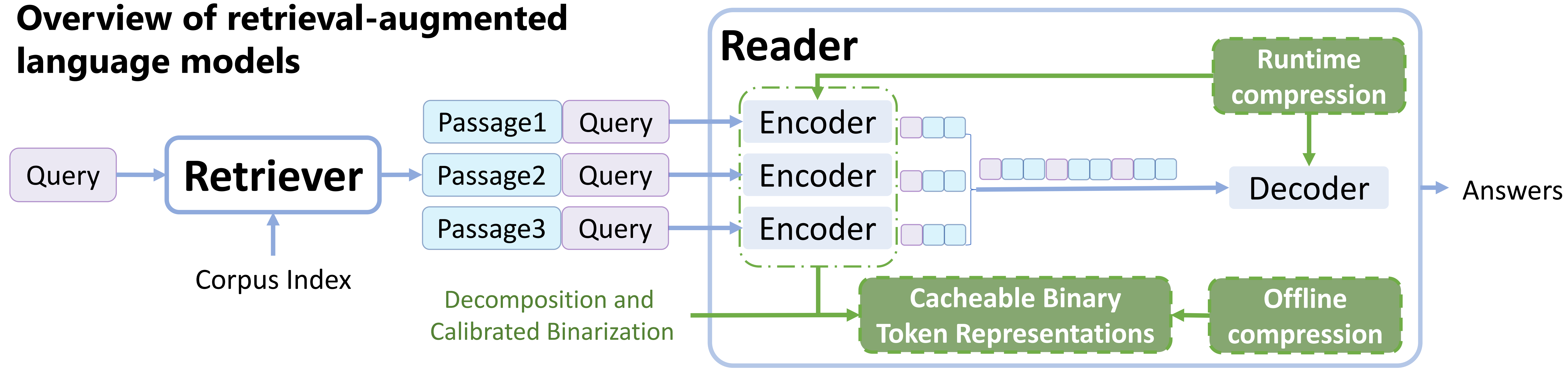
BTR: Binary Token Representations for Efficient Retrieval Augmented Language Models

Qingqing Cao, Sewon Min, Yizhong Wang, and Hannaneh Hajishirzi

UWNLP · UNIVERSITY of WASHINGTON

ICLR 2024 (spotlight)

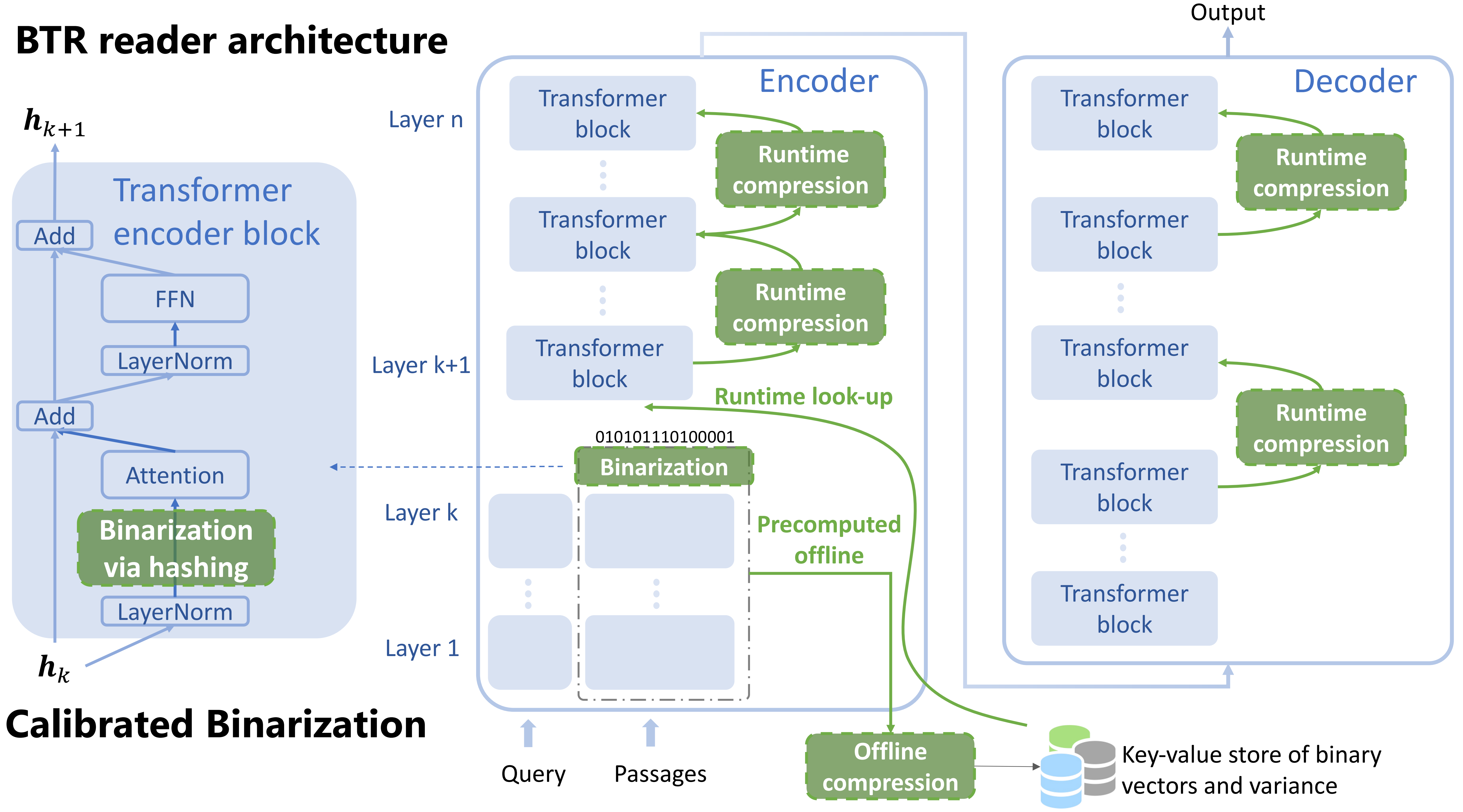
Overview of retrieval-augmented language models



- Retrieval-augmented models use a retrieve-and-read pipeline. The reader can be either an encoder or an encoder-decoder model.
- BTR creates cacheable binary representations for the passages via decomposition and calibrated binarization to speed up reader inference.
- BTR further reduces storage by offline compression and improves inference speed by runtime compression.

We create BTR: cacheable and calibrated binary token presentations that improve inference speed by >4x and reduce >100x storage for retrieval-augmented language models while maintaining knowledge-intensive NLP task performance.

BTR reader architecture



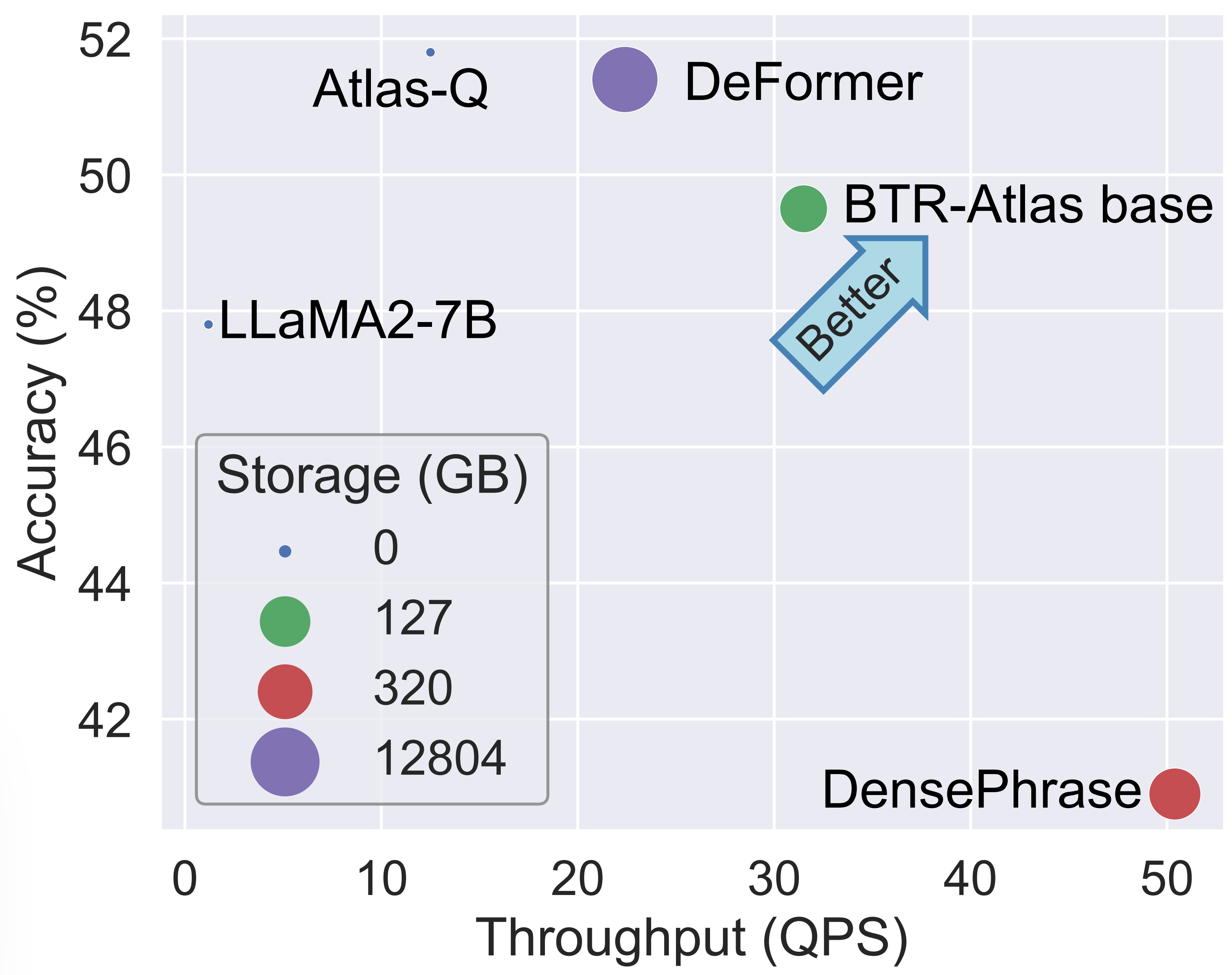
Calibrated Binarization

Offline and Runtime Compression

- Offline token compression reduces *context redundancy* so we do not store token representations every time it appears in a different context.
- Runtime token compression consists of intra-passage and cross-passage compression that remove similar information relevant to the query for different passages.

https://openreview.net/pdf?id=3TO3TtnOFl
(Or scan the left QR Code)

Major Results for NQ dataset


Accuracy (%)
Throughput (QPS)
Atlas-Q
DeFormer
BTR-Atlas base
LLaMA2-7B
Better
DensePhrase
Storage (GB)