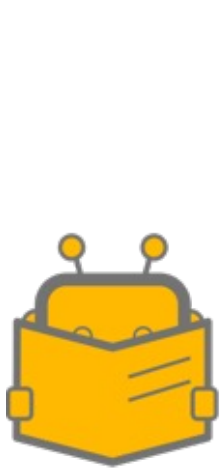
On The Adaptation of Unlimiformer for Decoder-Only Transformers



Microsoft

Project Turing

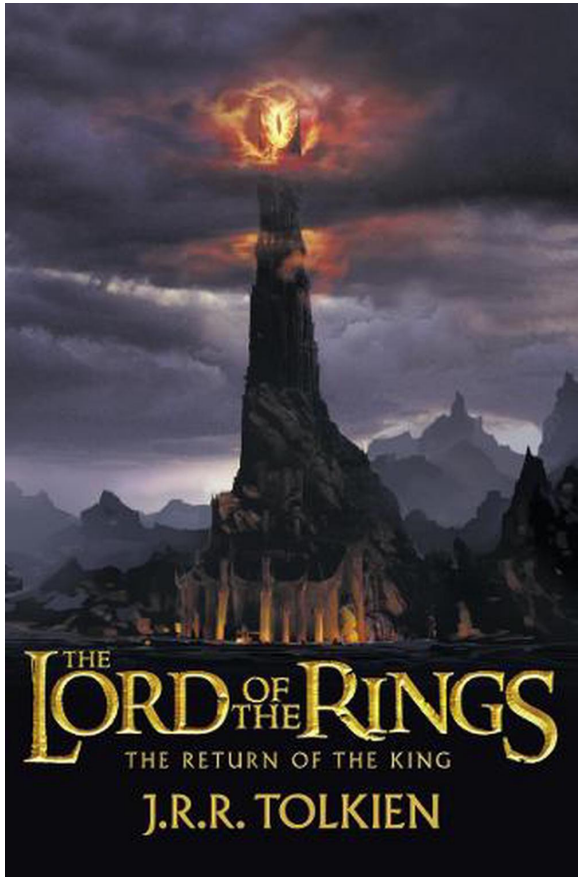
Making machines literate

Kian Ahrabian, Alon Benhaim, Barun Patra, Jay Pujara, Saksham Singhal, Xia Song

Information Sciences Institute

Motivations

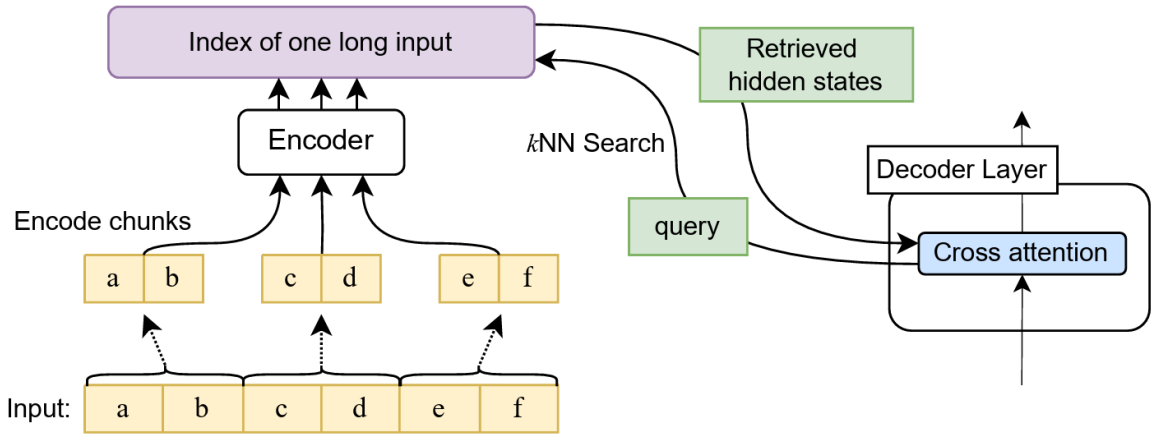
- One of the limitations of the transformers is that their inputs are bounded by their context length.
- For example, Llama-2 has a context length of 4k whereas average number of words in a book is between 70k and 100k tokens.
- Most of the recent prominent ideas could be clustered into four groups:
  - Extending positional embeddings through extrapolation/interpolation (e.g., xPOS, YaRN)
  - Introducing recurrence to the attention mechanism (e.g., Block-Recurrent Transformers, XL-NET)
  - Introducing sparsity to the attention mechanism (e.g., Block Sparse Attention)
  - Augmenting the attention mechanism with vector retrieval modules (e.g., Unlimiformer)



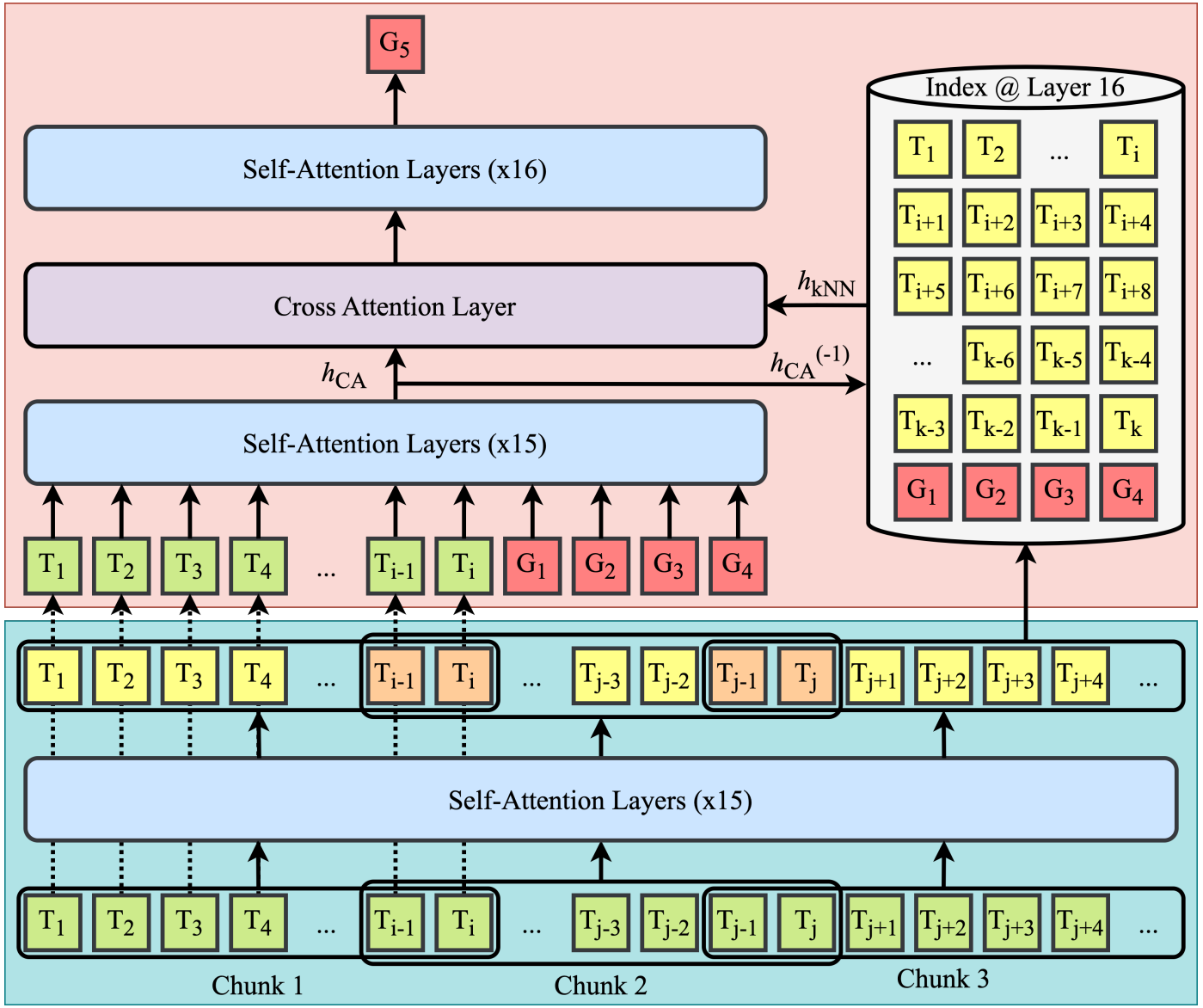
137.1k Words ~= 178.2k Tokens

Unlimiformer: Long-Range Transformers with Unlimited Length Input

- Split the input into overlapping chunks that the model can process.
- Encode each chunk separately and store the encoded hidden states in a k-NN datastore.
- At each decoder layer, retrieve the most important vectors from the datastore to approximate the full attention calculations.
- Authors show that this addition could work in both zero-shot and fine-tuned settings.
- In theory the model can process any input with no limitation on the length. But the authors didn't include any comparison with models that have longer context length.
- The goal of this project is to incorporate and adapt this module into the Turing LLM model and showcase its efficiency on at least one dataset when compared to a model with larger context length.

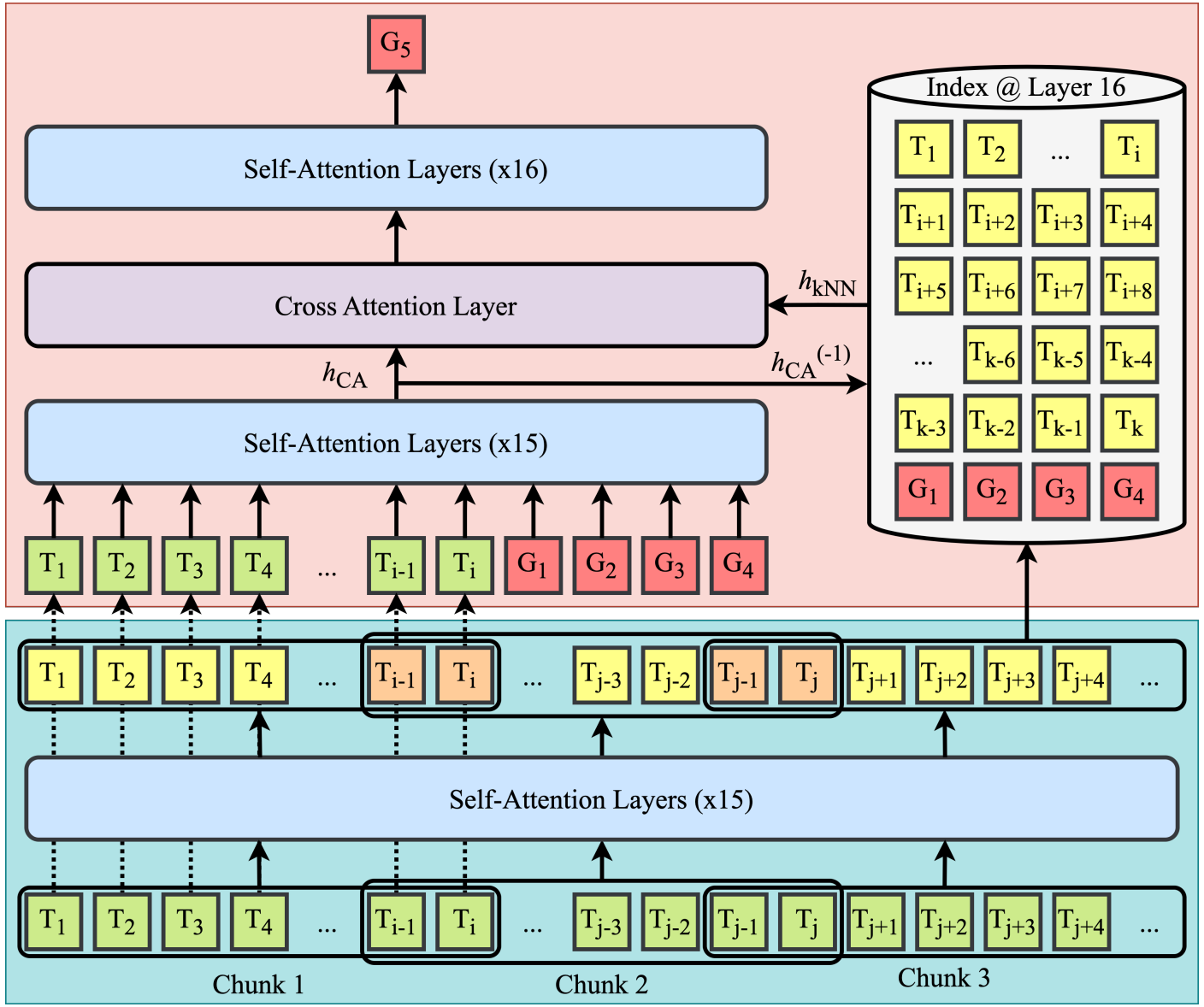


Architecture Overview

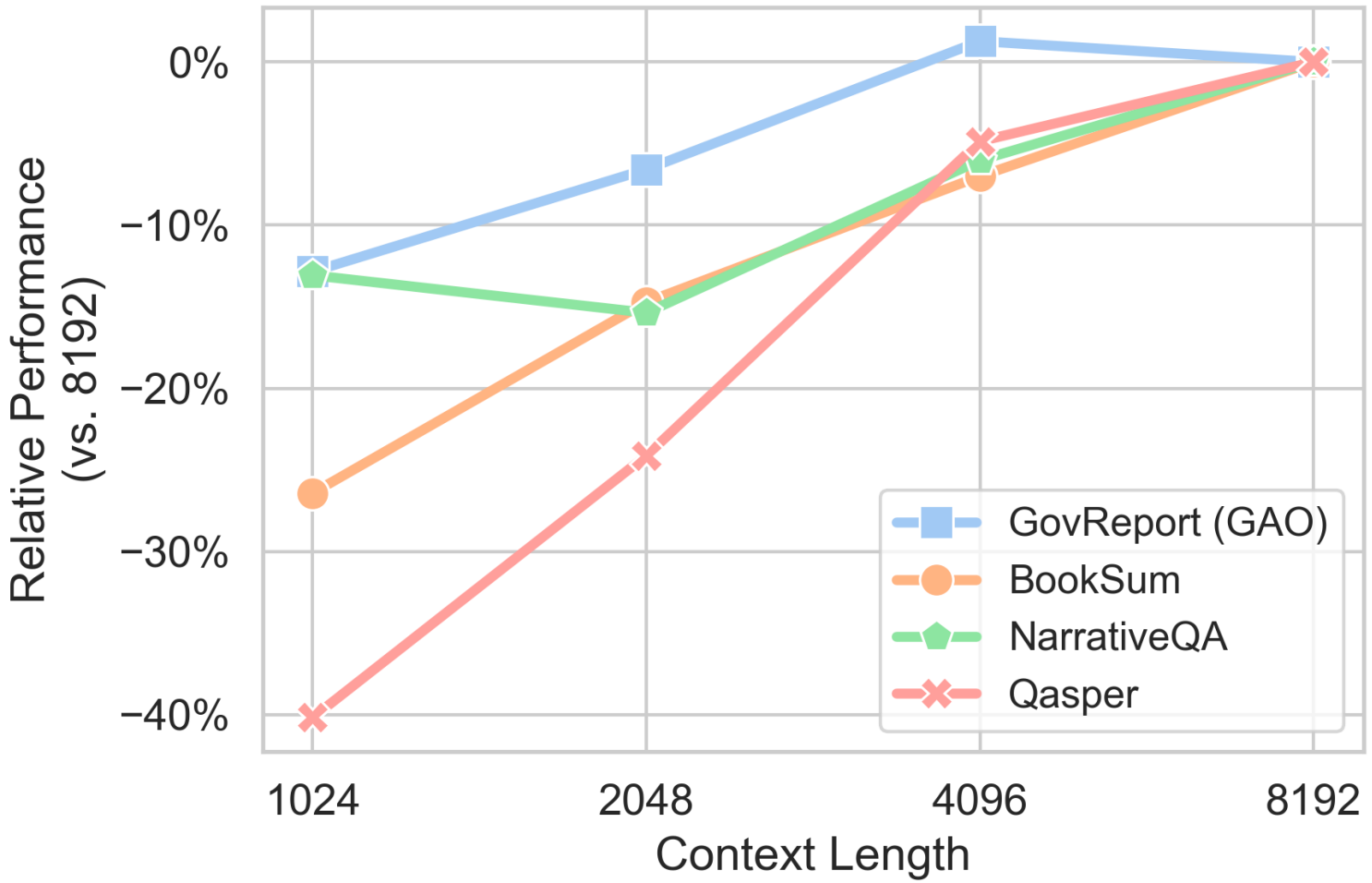


Architecture Changes

- Fused Cross-Attention
- Layer-wise kNN Indices
- Index Staleness Mitigator
- Updated Chunks Encoding

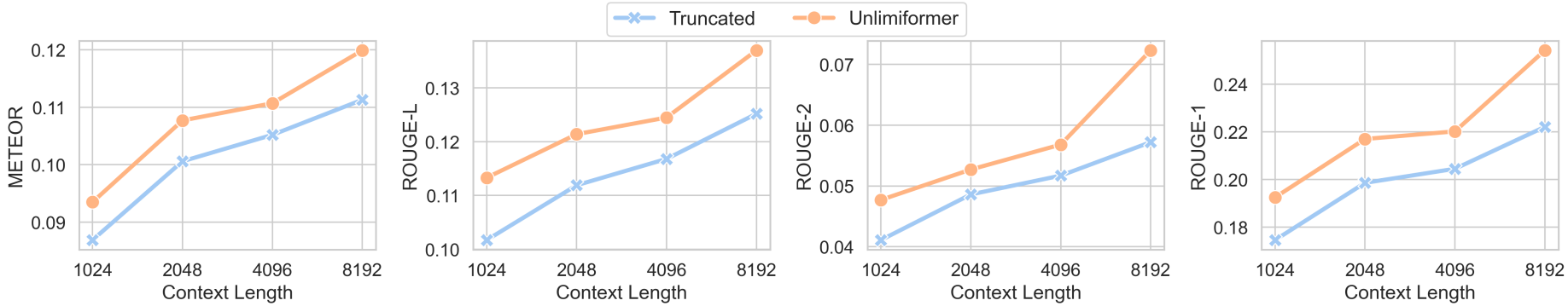


Long-Document Evaluation Dataset Explorations

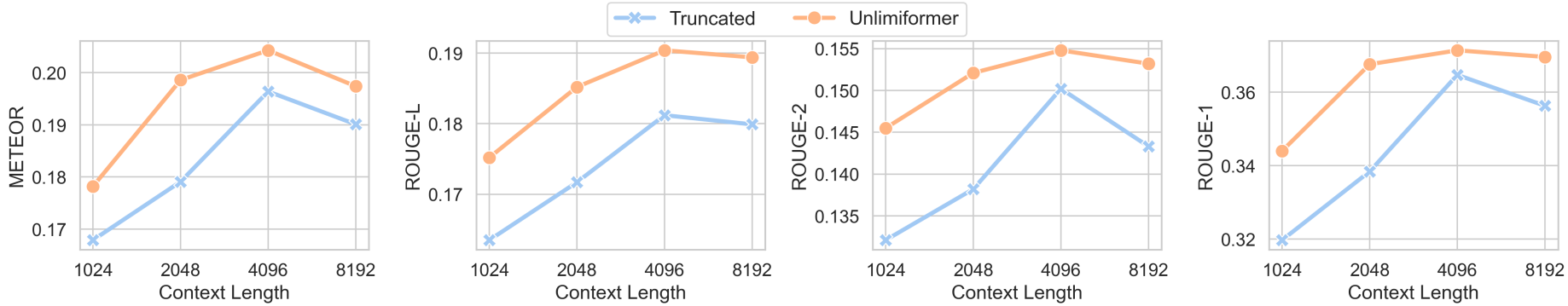


Summarization

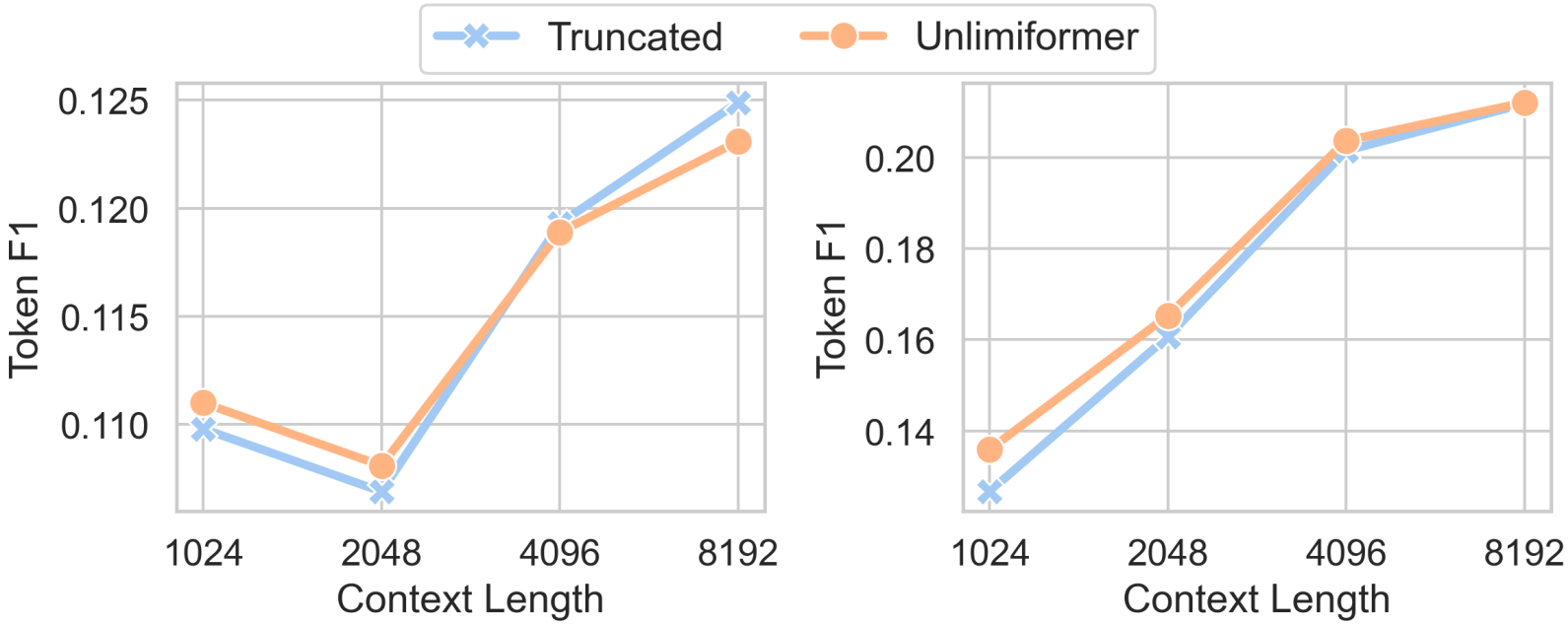
BookSum



GovReport (GAO)



Q&A



NarrativeQA

Qasper

Limitations

- Inherent overhead latency for calculating nearest neighbors while retrieving vectors.
  - Could be mitigated by using approximate indices.
- The improvements seem to be insignificant on the IFT model which requires further investigation.
- Query Bias

Takeaways and Future Directions

- The addition of the retrieval-augmented attention layers could help increase the upper-bound performance of the model, specially for summarization.
- Future experiments involving training and/or fine-tuning could help improve the performance of both IFT and summarization models.
- Future experiments could combine this approach (or other retrieval-based methods such as RETRO) with orthogonal approaches such as xPOS.

Thank You For Listening!