Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

Claas Beger, Cole Breen, Carl-Leander Henneking, Mihir Mishra, Max Whitton
Cornell University, Department of Computer Science, Ithaca NY, USA

Motivation & Problem Statement

Motivation

- Traditional language models rely on **parametric memory**, which struggles with knowledge updates and response transparency
- Knowledge-intensive tasks often require access to information not available in parametric memory
- **Non-parametric memory** allows for dynamic memory access but are not utilized by text generation models
- Retrieval-Augmented Generation enables a combination of parametric and non-parametric memory to improve factuality, interpretability, and adaptability

Problem Statement

- We re-implement and train RAG-Sequence with Fast Decoding to validate how document retrieval can improve text generation for **Question-Answering** and **Fact Verification**
- Given text corpus $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$ with latent retrieved documents z, query q , and target response , RAG-Sequence aims to maximize the following:

$$p_{\text{RAG-Sequence}}(y|x) \approx \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) p_\theta(y|x,z) = \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) \prod_i^N p_\theta(y_i|x,z,y_{1:i-1})$$

Background

RAG Architecture

- Dense Passage Retrieval (DPR) with BART generator
- Two RAG variants:
  - RAG-Sequence: uses a single retrieved document for the full output generation

$$p_{\text{RAG-Sequence}}(y|x) \approx \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) p_\theta(y|x,z) = \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) \prod_i^N p_\theta(y_i|x,z,y_{1:i-1})$$

  - RAG-Token: each generated token uses a (possibly) different retrieved document

$$p_{\text{RAG-Token}}(y|x) \approx \prod_i^N \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) p_\theta(y_i|x,z,y_{1:i-1})$$

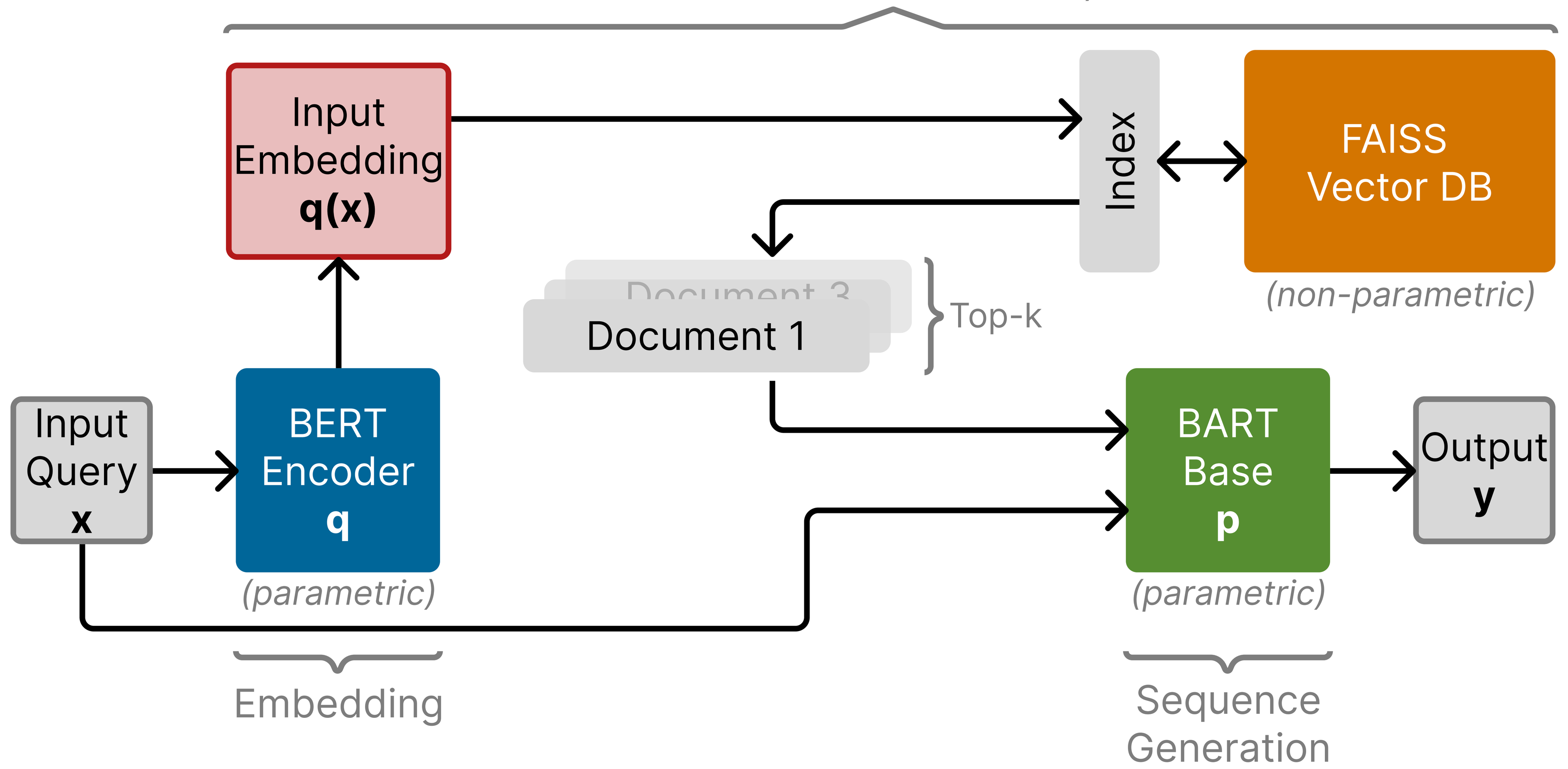
- Two decoding strategies (method of generating output sequences):
  - Thorough Decoding: uses a combination of beam search and full re-calculation of output probabilities on candidate outputs
  - Fast Decoding: uses beam search per candidate output, approximating the marginal likelihood calculation; faster than thorough decoding

Datasets

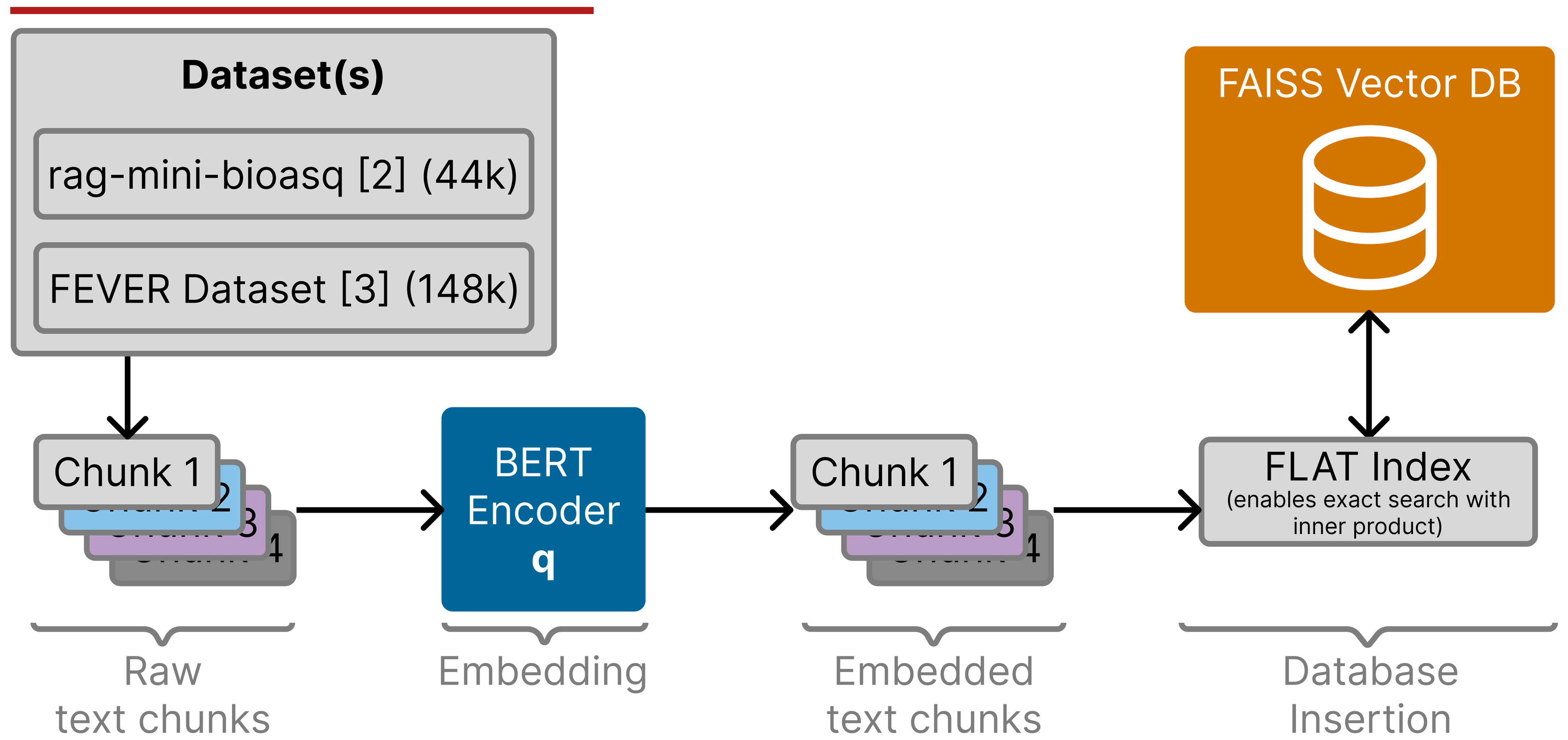
- rag-mini-bioasq [2]: a biomedical Question-Answering dataset from HuggingFace
- FEVER [3]: a fact verification dataset with claims labeled as "Supported", "Refuted", or "Not Enough Info"

Methodology

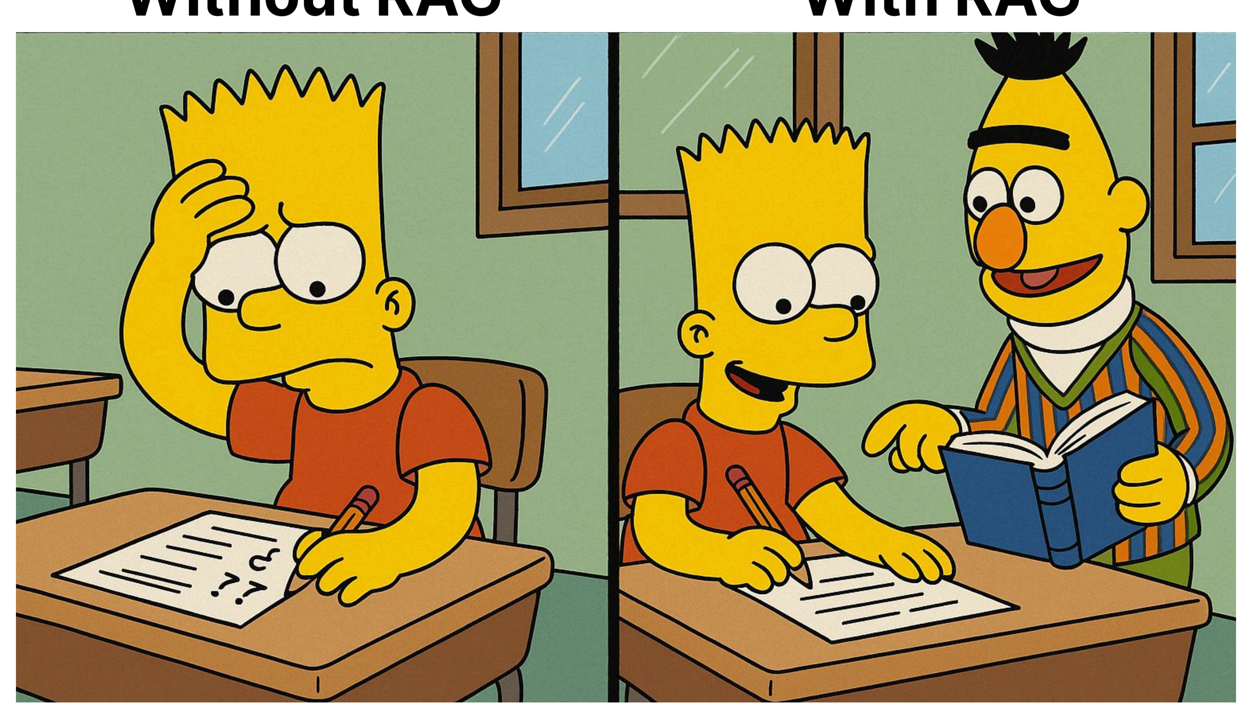
Overview



Database Initialization



Without RAG **With RAG**



Results

QA Setup and Evaluation

- Test split was 20% of the questions in rag-mini-bioasq (about 800 QA pairs)
- Generated answer was compared to "ground truth" answer for baseline BART and trained retrieval augmented BART
- At test time, K (# of documents retrieved) is 1
- **BLEU-1** is % overlap between generated answer and ground truth answer
- **ROGUE-L** is the longest common subsequence between generated answer and ground truth answer

mini-bioasq Results

Approach	Avg BLEU-1	Avg ROUGE-L
Baseline	0.1086	0.2225
RAG	0.4355	0.3860

Fact Verification Setup and Evaluation

- **Macro Precision** is the average % of correct positive predictions across all classes
- **Macro Recall** is the average % of actual positives correctly predicted across all classes
- **Macro F1** is the average harmonic mean of precision and recall across all classes
- At test time, K (# of documents retrieved) is 1

FEVER Results

Approach	Accuracy	Macro-Precision	Macro-Recall	Macro-F1
Baseline	0.2380	0.0793	0.3333	0.1282
RAG	0.6562	0.6626	0.6354	0.6372

Conclusion

- As in the original paper, we displayed leveraging a hybrid fine-tuned generative model, **combining both parametric memory and non-parametric memory**, in order to outperform a baseline parametric model
- We demonstrated that our **RAG model achieves strong performance** on a question-answer task in the biomedical domain, which is characterized by highly specialized terminology, concepts, and questions
  - This contrasts the original paper which tested open domain knowledge
- We demonstrated that our RAG model achieves **higher accuracy and performance** than baseline methods when fine-tuned and evaluated on the FEVER fact verification classification task
- **Limitations** might include **dataset size and generative model complexity**
- Future work may include delving into optimal hyper-parameters at train time (including learning rate and k-value) and potentially updating the document encoder and embeddings at train time

References

[1] Lewis, Patrick, et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." ArXiv, 2020, https://arxiv.org/abs/2005.11401.
[2] "Rag-Datasets/Rag-Mini-Bioasq · Datasets at Hugging Face." Rag-Datasets/Rag-Mini-Bioasq, Hugging Face, huggingface.co/datasets/rag-datasets/rag-mini-bioasq.
[3] Thorne, James, et al. "FEVER: A Large-scale Dataset for Fact Extraction and VERification." ArXiv, 2018, https://arxiv.org/abs/1803.05355.