॥ सा विद्या या विमुक्तये ॥

भारतीय प्रौद्योगिकी संस्थान धारवाड़

Indian Institute of Technology Dharwad

Dense Passage Retrieval for Open-Domain Question Answering

Rishith	210010002
Sivamohan	210020025
Sreenivasu	210020001
Akash Anand	210030002

Introduction

PROBLEM STATEMENT:

- Traditional open-domain QA systems are complex, involving multiple components for question retrieval and answer extraction from vast document collections.

- This project aims to address the challenge by proposing a Dense Passage Retriever (DPR), trained without additional pre training, using pairs of questions and passages.

- The QA process into a two-stage framework, where a context retriever selects relevant passages, followed by a machine reader to extract answers.

- Leveraging BERT and a dual-encoder architecture, the approach optimizes embeddings to maximize inner products between question and passage vectors, achieving competitive or superior results in the open-retrieval setting.

Research Objectives:

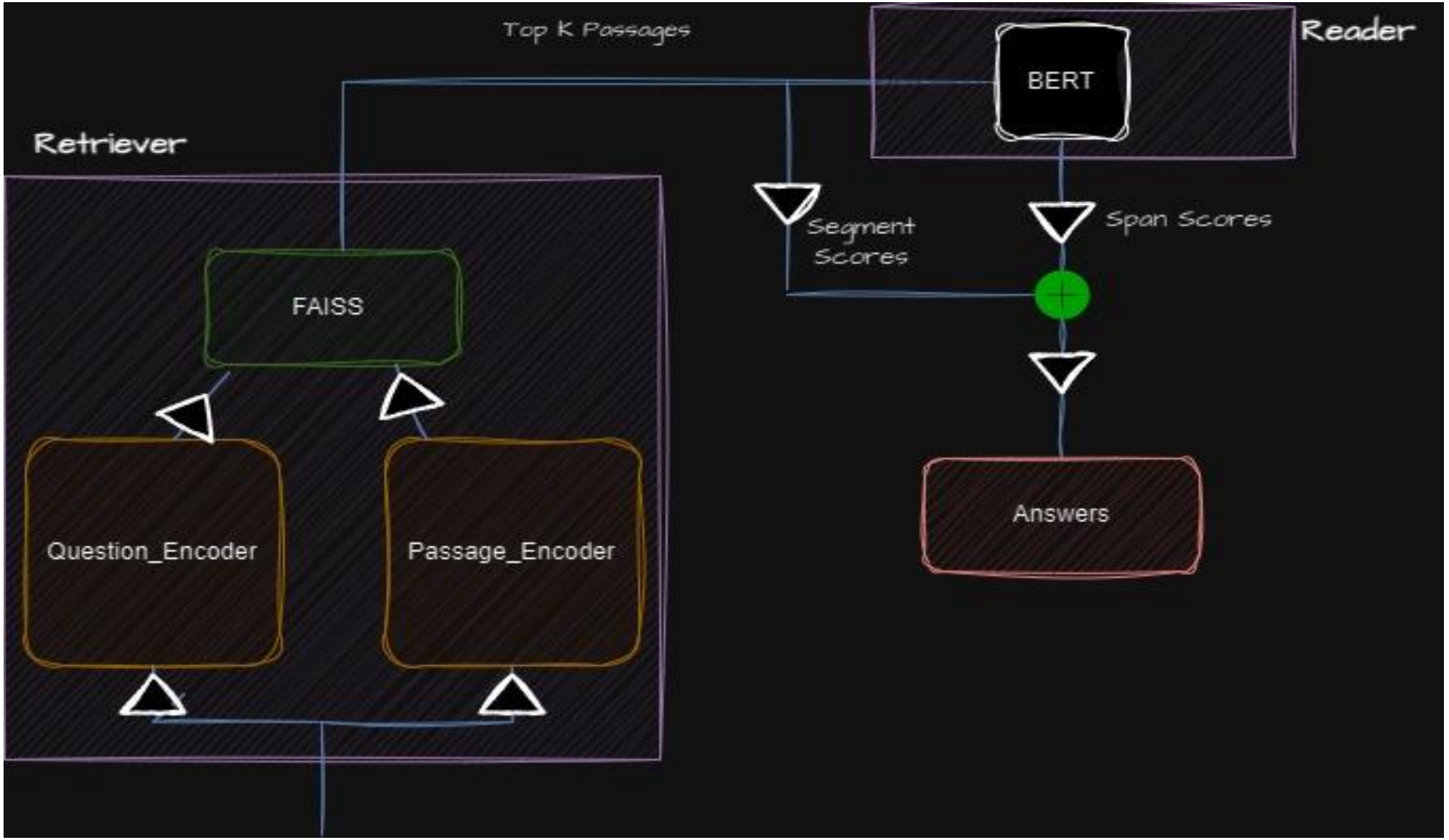
- Implement various dense retrieval techniques to surpass traditional sparse models like TF-IDF and BM25, thereby improving the quality of open-domain question answering systems.

- Develop scalable and efficient strategies for dense retrieval methods to overcome computational constraints in large-scale QA systems with retriver top k passages are obtained and with reader part answer is obtained from passages.

- Assess the ability of dense encodings to capture semantic information, including synonyms, paraphrases, to enhance the precision and relevance of retrieved answers.

- Evaluate the effectiveness of dense encodings in complex context matching, particularly in scenarios with limited keyword overlap, to achieve more accurate retrieval and better match user queries with relevant information.

Architecture



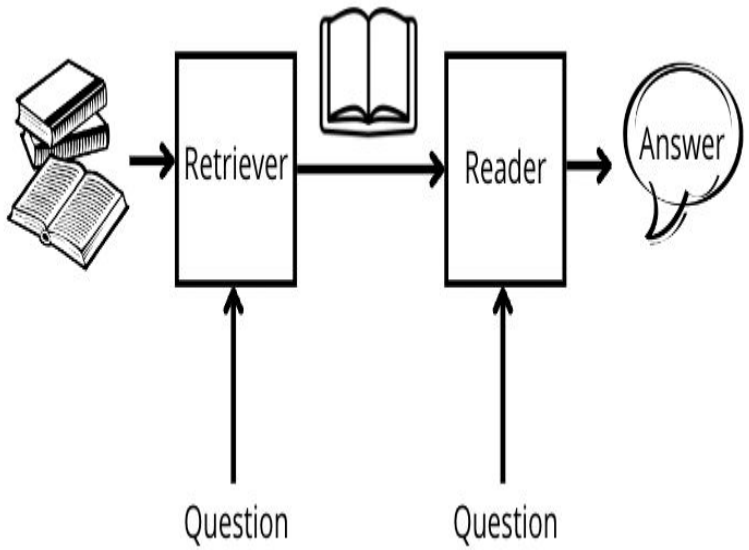
METHODOLOGY

Dense Passage Retriever

- Utilizes dense encoders to map questions and passages into a low-dimensional continuous space for efficient retrieval.Aims to index a large collection of passages (M) into a low-dimensional space, enabling quick retrieval of relevant passages.

- **Indexing Passages**:Utilizes a dense encoder (EP()) to map each passage into a d-dimensional vector.Builds an index for all passages, even with a large collection (M).

- **Question Encoding**:Applies a specific encoder (EQ()) to map input questions to a d-dimensional vector, capturing semantic meaning.

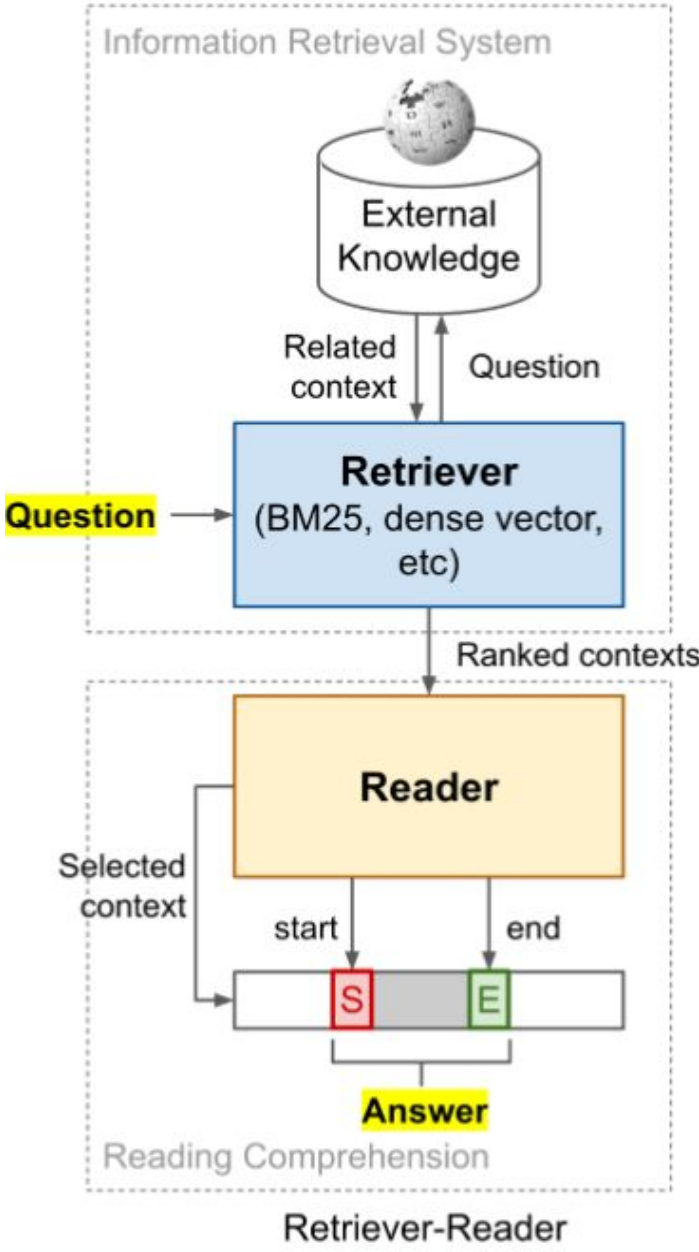


Dense Passage Retriever

- **Retrieval Process:**Computes similarity between question vector and passage vectors using dot product operation.Utilizes inner product function for simplicity and effectiveness in similarity computation.

- **Encoders:**Employs two independent BERT networks (base, uncased) as encoders for questions and passages.Outputs representations at the [CLS] token resulting in d = 768-dimensional vectors.

- **Inference**:Applies passage encoder (EP) offline to all passages and indexes them using FAISS for efficient retrieval.Retrieves top k passages with embeddings closest to the question embedding (vq = EQ(q)).
- **Reader Process** : DPR may include a reader module responsible for processing retrieved passages to generate final answers to the input questions.



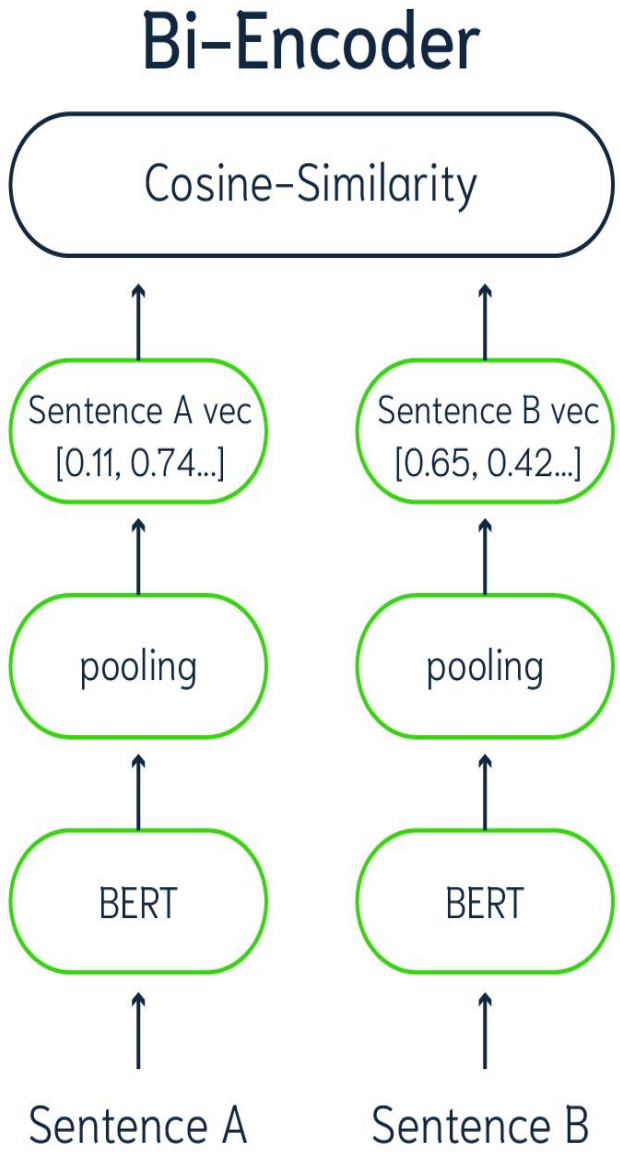
Retriever-Reader

Bi-Encoder:

- Employs two BERT models for encoding questions and passages.Utilizes bidirectional attention mechanisms of BERT to capture contextual information.

- **Dataset Preparation**:Custom QADataset class handles pairs of questions and passages, tokenizing inputs using BERT tokenizer.

- **Model Definition**:DualEncoderModel architecture consist of two BERT encoders for questions and passages, extracting pooled output embeddings of [CLS] token.



Bi-Encoder:

- **Triplet Loss Function**:Computes triplet loss between anchor, positive, and negative embeddings to encourage model to minimize distance between anchor and positive embeddings while maximizing distance between anchor and negative embeddings.

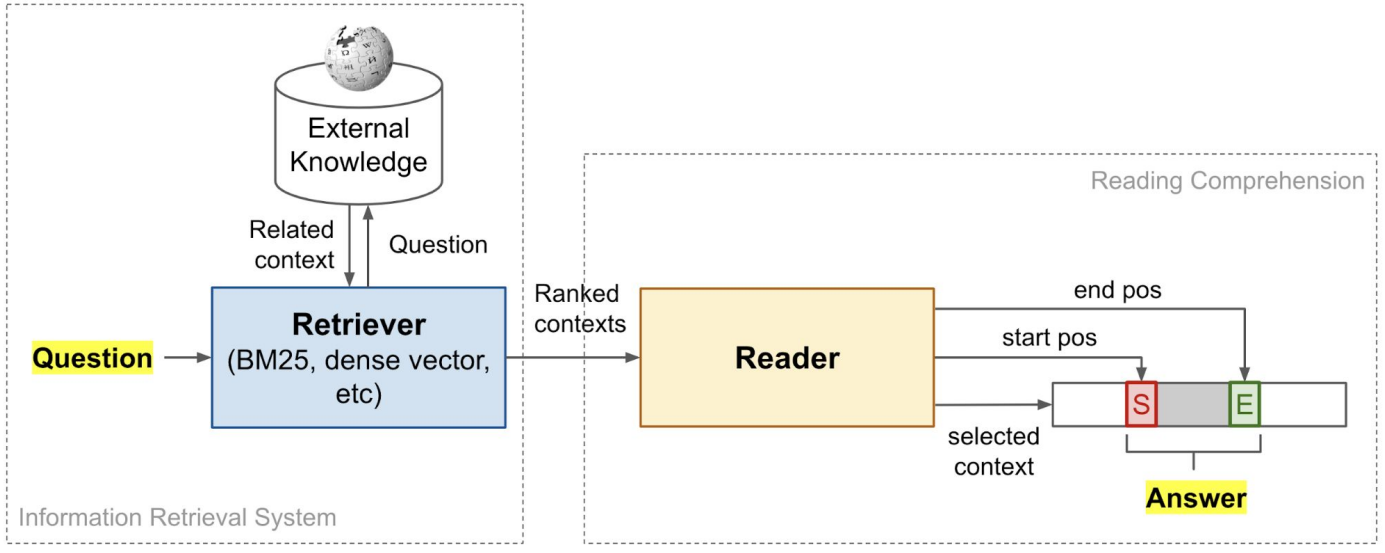
- **Model Initialization:**Initializes BERT models for question and passage encoding using bert-base-uncased pre-trained weights.

- **Training Setup**:Initializes Adam optimizer with learning rate of 1e-5, sets number of epochs and margin for triplet loss.Defines dataset using QADataset class and creates data loader for batch iteration during training.

- **Training Loop**:Iterates over each epoch and batch of data, computes question and passage embeddings using the model.Generates negative embeddings for triplet loss, computes loss, and updates model parameters.Monitors progress using tqdm, displaying loss for each batch.

Reader:

- **Tokenization**:Utilizes BERT tokenizer (tokenizer.encode) to convert question and context text into numerical token IDs suitable for model input.
- **Chunking**: Divides the context into chunks of a specified size (chunk size) to handle large contexts efficiently.
- **Model Prediction:**For each chunk of the context combined with the question, invokes the model(input ids) to obtain predictions.
- **Output**: Model outputs include start logits and end logits, representing scores for potential start and end positions of the answer within the context.



Reader:

Top Answer Extraction:

- **Identification:**Identifies indices corresponding to the top N scoring positions from the start logits and end logits (possible starts and possible ends).
- **Extraction**: Extracts answer spans (start and end positions) corresponding to these top scoring indices.

Answer Post-Processing

- **Answer Conversion**: Converts identified answer spans back from token IDs to a string representation (tokenizer.convert tokens to string).
- **Irrelevant Answer Filtering**: Filters extracted answers to retain only those containing relevant information related to the question (filter irrelevant answers).

RESULTS

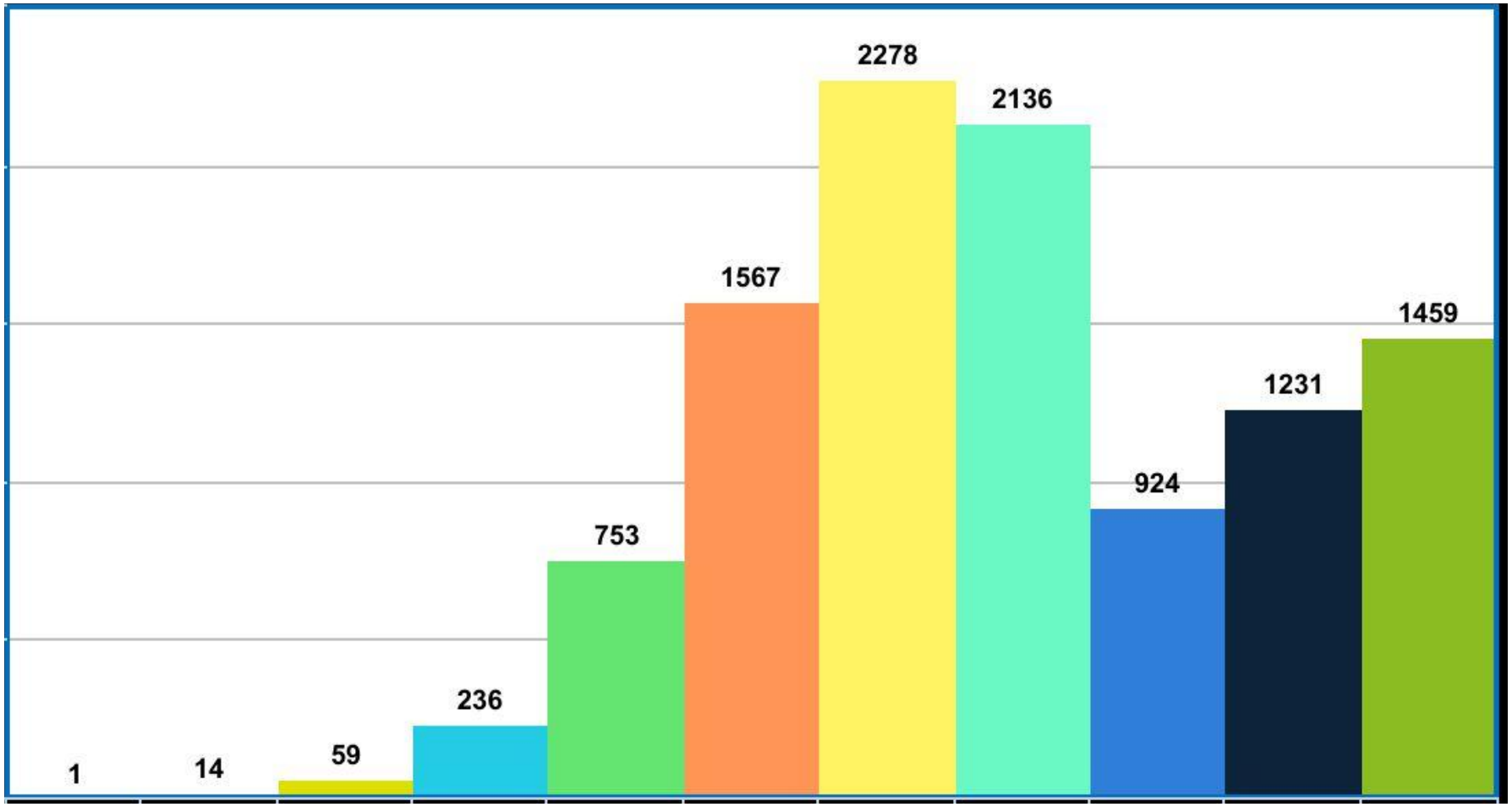
Question: When will the sydney light rail be completed?
Top Passage Context: The CBD and South East Light Rail is a future Australian light rail line in Sydney, New South Wales, running from Circular Quay at the northern end of the central business district to the south-eastern suburbs of Randwick and Kingsford. The line will be part of Sydne y's light rail network. Major construction commenced in October 2015. The project is being manag ed by the New South Wales Government's transport authority, Transport for NSW. Construction, ope ration and maintenance of the line is contracted to the ALTRAC Light Rail consortium.

Ground Truth Passage: The CBD and South East Light Rail is a future Australian light rail line i n Sydney, New South Wales, running from Circular Quay at the northern end of the central busines s district to the south-eastern suburbs of Randwick and Kingsford. The line will be part of Sydn ey's light rail network. Major construction commenced in October 2015. The project is being mana ged by the New South Wales Government's transport authority, Transport for NSW. Construction, op eration and maintenance of the line is contracted to the ALTRAC Light Rail consortium.

Question: When did christianity become official religion of rome?
Text: Nicene Christianity became the state church of the Roman Empire with the Edict of Thessalonica in 380 AD, when Emperor Theodosius I made it the Empire's sole authorized religion.[1][2] The Eastern Orthodox Church, Oriental Orthodoxy, and the Catholic Church each claim to be the historical continuation of this church in its original form, but do not identify with it in the caesaropapist form that it took later. Unlike Constantine I, who with the Edict of Milan of 313 AD had established tolerance for Christianity without placing it above other religions[3] and whose involvement in matters of the Christian faith extended to convoking councils of bishops who were to determine doctrine and to presiding at their meetings, but not to determining doctrine himself,[4] Theodosius established a single Christian doctrine (specified as that professed by Pope Damasus I of Rome and Pope Peter II of Alexandria) as the Empire's official religion.

Answers:
1. theodosius established a single christian doctrine (specified as that professed by pope damsus i of rome and pope peter ii of alexandria) as the empire ' s official religion
2. nicene christianity became the state church of the roman empire with the edict of thessaloni a in 380 ad
3. when did christianity become official religion of rome? diocletianic persecution of 303 – 31 and the donatist controversy that arose in consequence, constantine had convened councils of christian bishops to define the orthodoxy, or " correct teaching ", of the christian faith, expand ng on earlier christian councils.

```
Number of questions with sim_score_ground_truth between 0.0 and 0.1: 1
Number of questions with sim_score_ground_truth between 0.1 and 0.2: 14
Number of questions with sim_score_ground_truth between 0.2 and 0.3: 59
Number of questions with sim_score_ground_truth between 0.3 and 0.4: 236
Number of questions with sim_score_ground_truth between 0.4 and 0.5: 753
Number of questions with sim_score_ground_truth between 0.5 and 0.6: 1567
Number of questions with sim_score_ground_truth between 0.6 and 0.7: 2278
Number of questions with sim_score_ground_truth between 0.7 and 0.8: 2136
Number of questions with sim_score_ground_truth between 0.8 and 0.9: 924
Number of questions with sim_score_ground_truth between 0.9 and 1.0: 1231
Number of questions with sim_score_ground_truth between 1.0 and 1.1: 1459
```


1
14
59
236
753
1567
2278
2136
924
1231
1459

Conclusion

- Our experimentation results show promising performance trends comparable to those observed by the original author. Both our retriever and reader models exhibit potential for further improvement through additional training and fine-tuning.

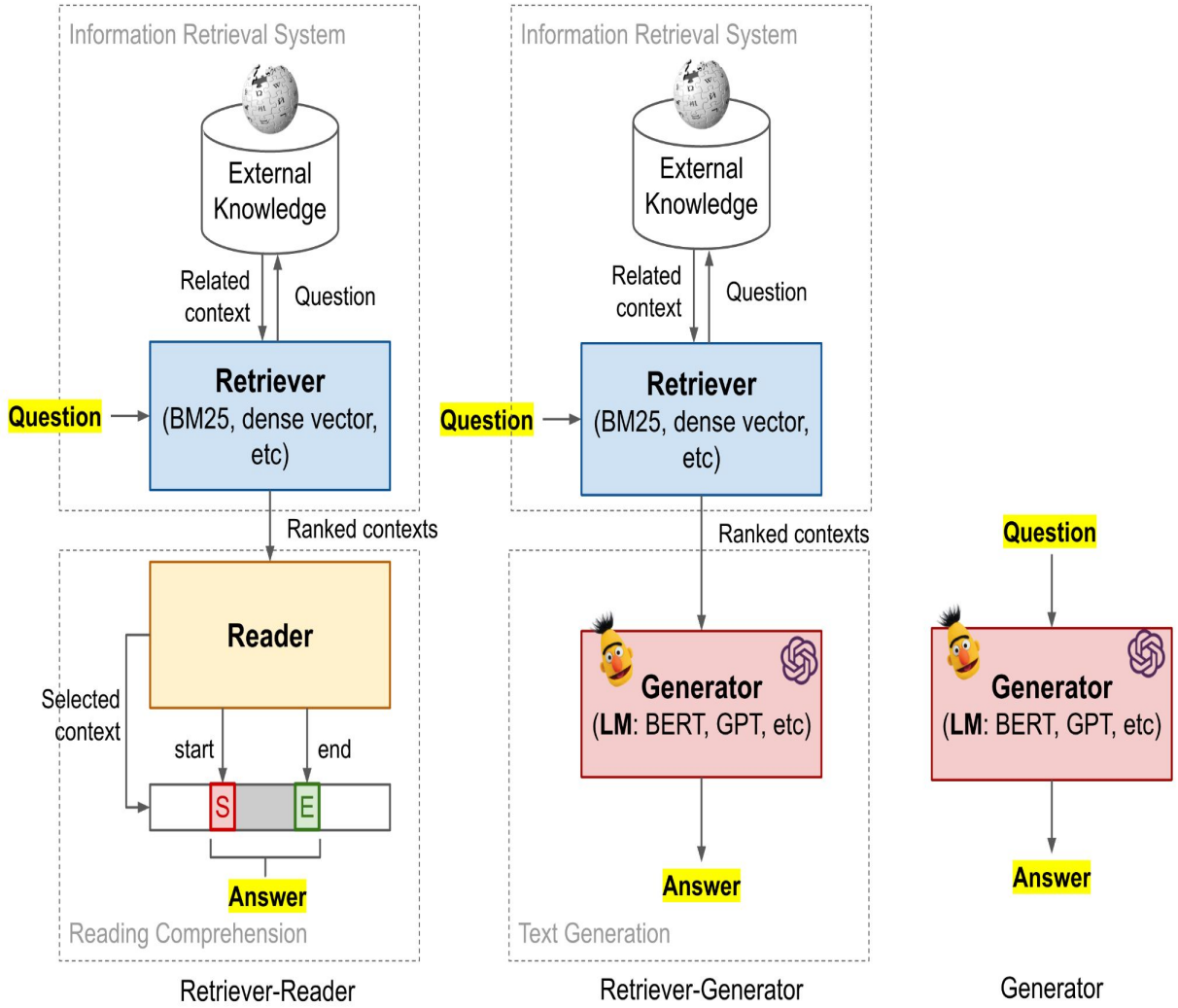
- **Retriever Model Performance**:Our retriever model successfully recommends contexts closely resembling the ground truth passages, indicating its effective ability to retrieve relevant information.Further optimization could enhance its accuracy, leading to even better retrieval performance.

- **Reader Model Performance:**The reader model demonstrates its capability to extract accurate answers from the provided contexts.Extracted answers align well with the corresponding questions, highlighting the reader's understanding and comprehension of the input data.

Future Scope

- **Integration of Generative Models:** Incorporate models like GPT alongside the reader component to generate nuanced and contextually rich answers, enhancing response quality.

- **Exploring RAG Models for Enhanced Retrieval:** Investigate advanced retrieval models such as RAG to augment the retriever's capabilities, leveraging large-scale language models for improved relevance and diversity of retrieved passages.



Future Scope

- **Combining Techniques for Comprehensive QA:** Explore hybrid approaches that combine generative and retrieval-based methods to address complex query scenarios and provide more comprehensive answers.

- **Fine-tuning with Domain-Specific Data**: Consider fine-tuning the question-answering system with domain-specific data to improve its performance on specific tasks or domains. Fine-tuning on relevant datasets can enhance the system's understanding of domain-specific terminology and nuances, leading to more accurate and tailored responses.

Thank You