

IR Meets LLM

Group 04


Members: Stephan Bartl, Dorian Percic

Repository: <https://github.com/stephba/IR-meets-LLM>

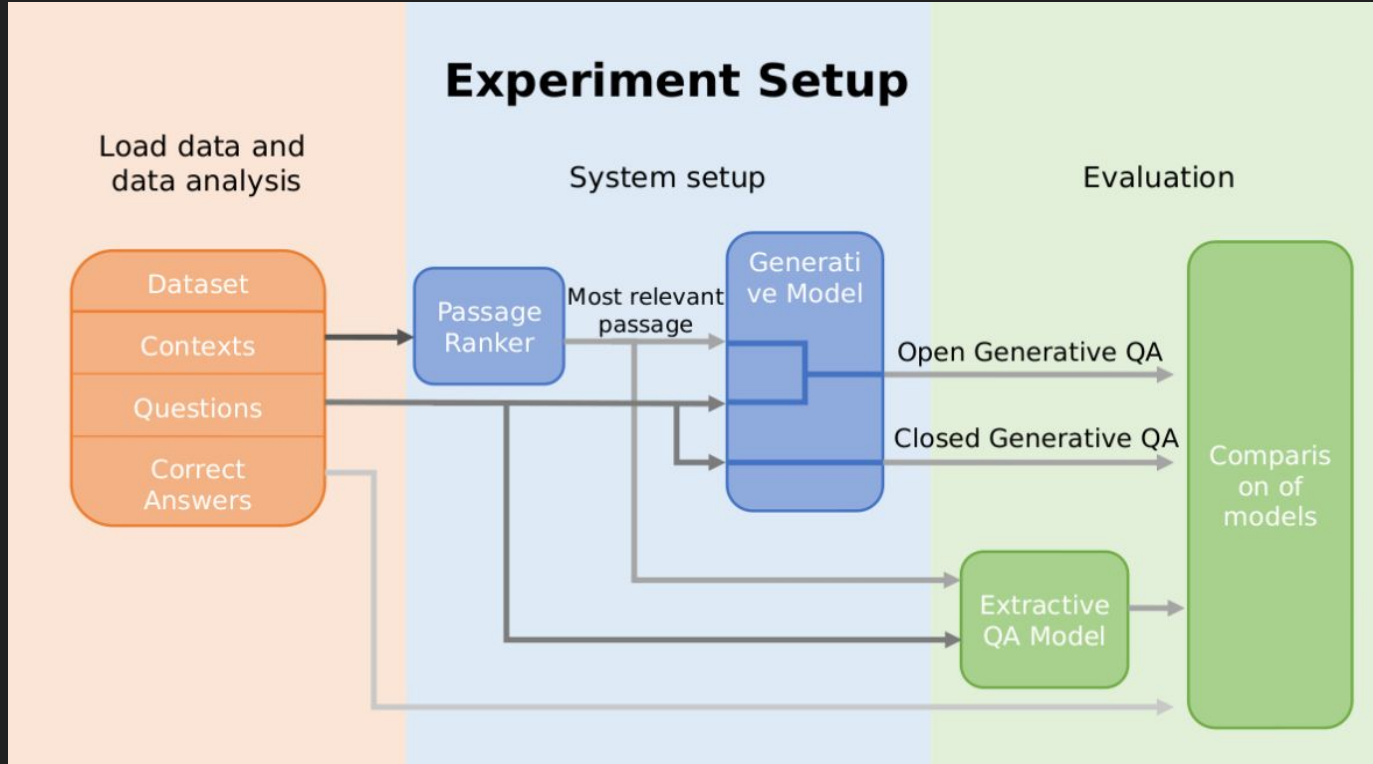
Roles: Stephan: Planning of general architecture, project implementation

Dorian: Basic data analysis, presentation

Introduction

- R&D of Q&A models has risen. 
- 3 Main Categories:
 - Extractive Model
 - Open Generative Model
 - Closed Generative Model
- IR: Passage ranking using a *Bi-Encoder* to first retrieve context.
- Analysis of performance of each model based on different metrics.
- Analysis of strengths and weaknesses of each model.

Introduction



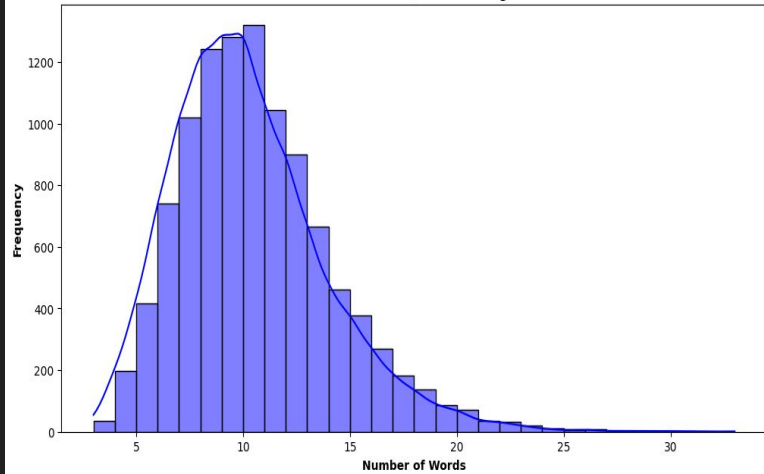
Data and Methods

- Data for training and evaluation: *SQuAD* dataset.
 - 100 000 question/answers with given context.
 - Split into test (evaluation) and train set (model training).
- Open and closed generative model: *OPT* (Facebook).
 - 1.3 Billion parameters.
- Extractive model: *deepset/roberta-base-squad2* from *Hugging Face* 🙌

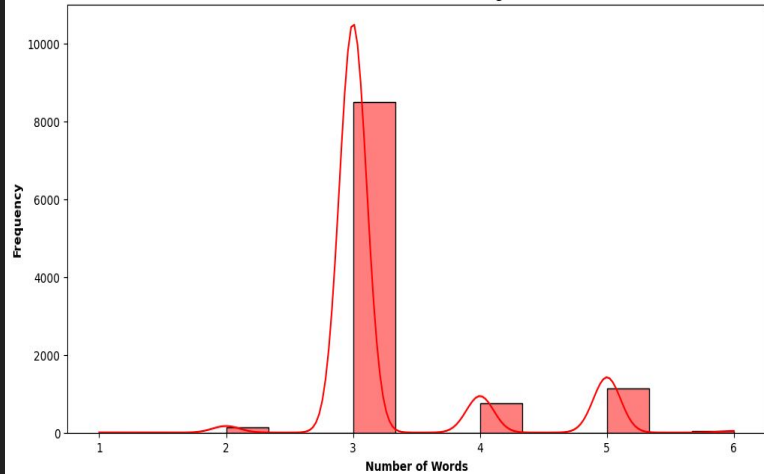
- Fine tuned using *SQuAD 2.0*.



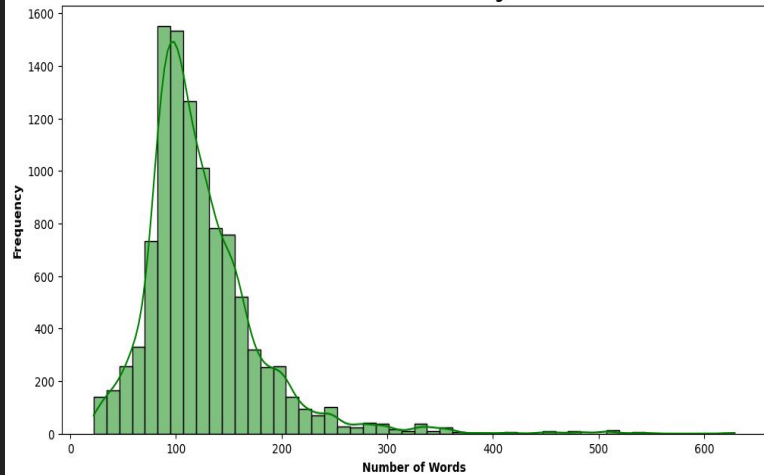
Distribution of Question Lengths



Distribution of Answer Lengths



Distribution of Context Lengths



Results - Evaluation Metric

- Initial Evaluation Ideas:
 - Exact match: In Generative Model score always 0.
 - *BLEU* method: Sentence length ≥ 4 .
- Ultimately evaluation through precision and recall:
 - Precision: $\frac{|\text{Correct words}|}{\text{Length of model answer}}$
 - Recall: $\frac{|\text{Correct words}|}{\text{Length of correct answer}}$

Results - Model Ranking

- Results in both of the metrics:

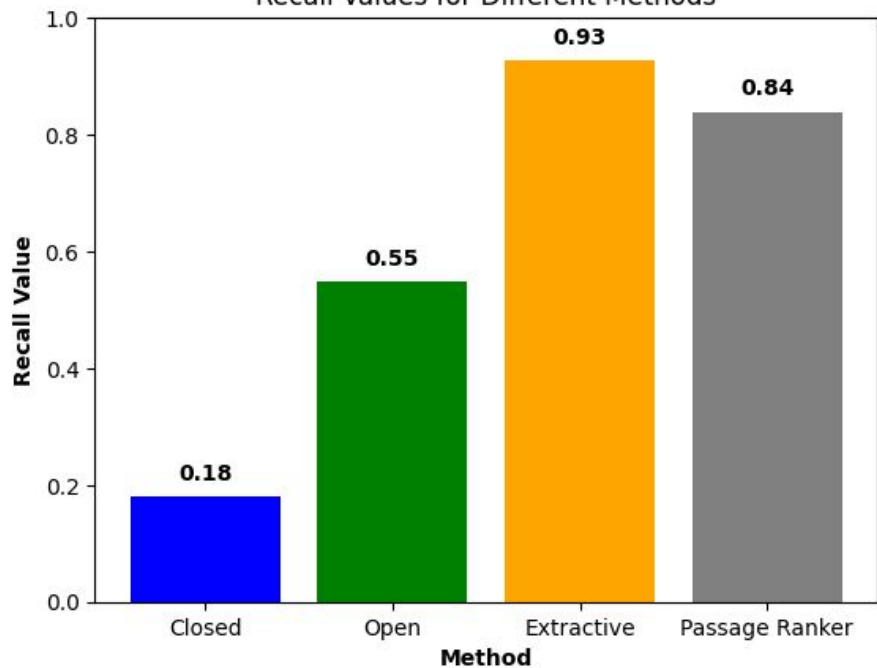
- ①. Extractive Model
- ②. Open Generative Model
- ③. Closed Generative Model

- Answer generation speed:

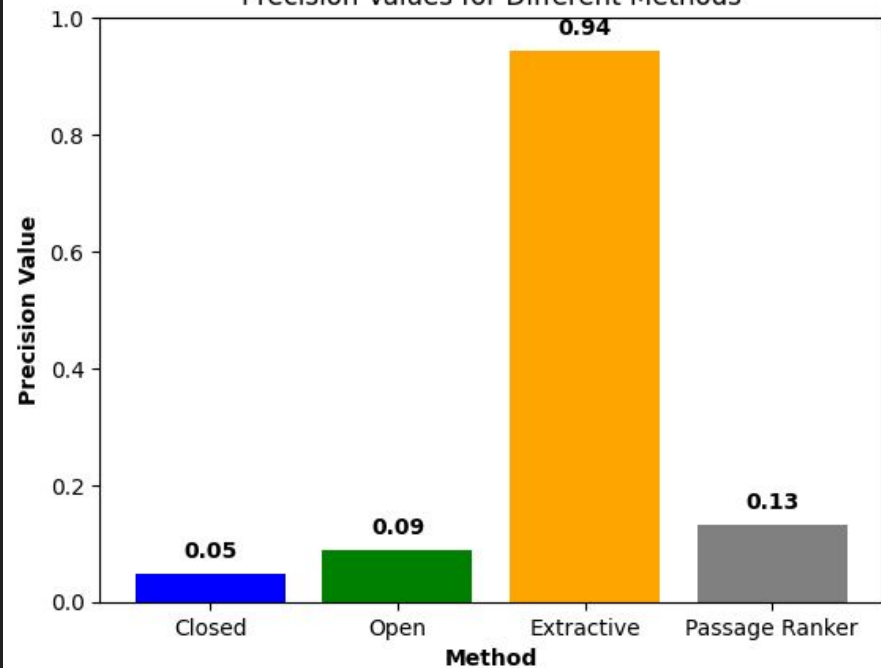
- ①. Extractive Model
- ②. Closed Generative
- ③. Open Generative

Results - Plots

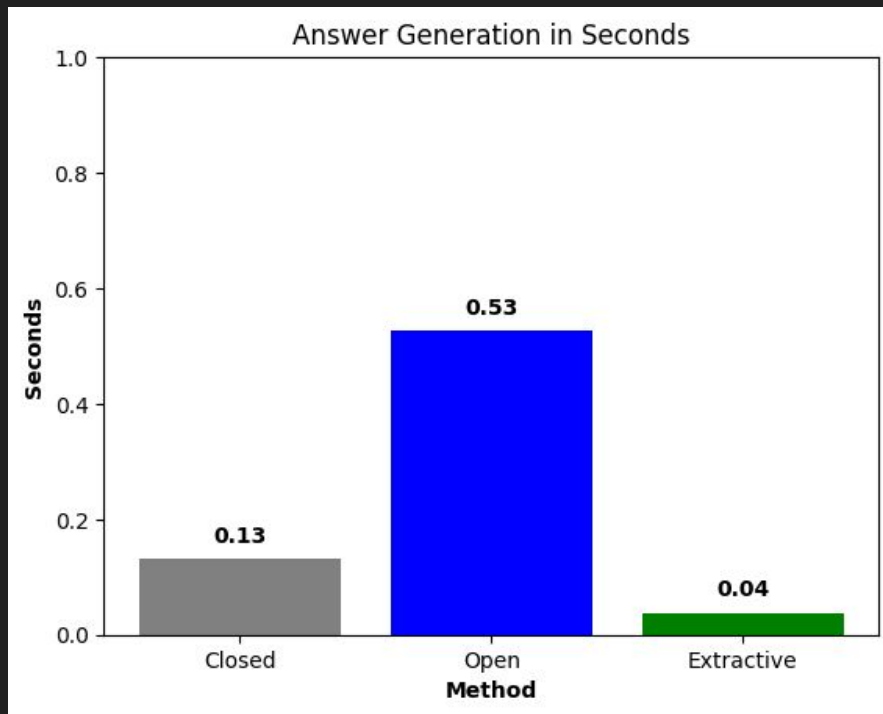
Recall Values for Different Methods



Precision Values for Different Methods

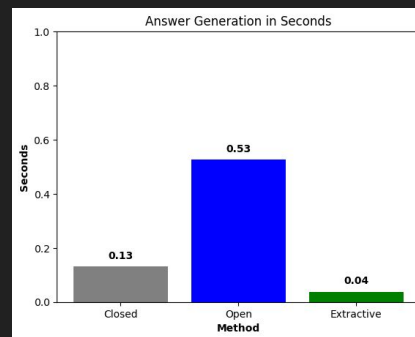
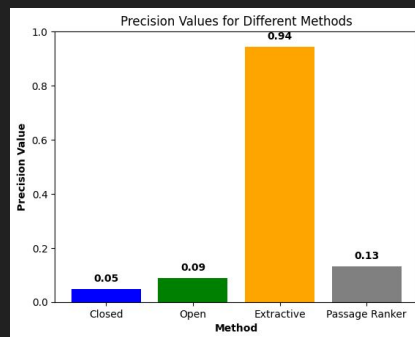
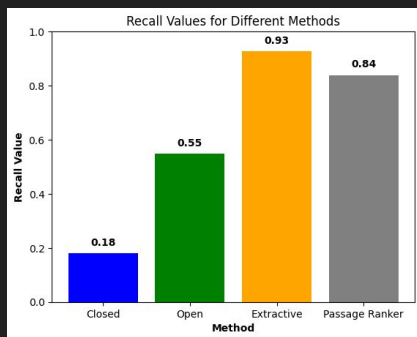


Results - Plots



Results - Analysis

- Baseline model performs the best, as it is fine tuned using *SQuAD 2.0*.
- OGM 3x more correct words than CGM (Recall).
- OGM precision score low, still 2x more compact than CGM:
 - + Whole and longer sentence, often with additional information.
 - Read whole sentence to get information.
- Passage rankers high recall score indicates its functionality.



Conclusion

- Positive impact of context as additional input to generative model.
- Faster answer generation through Passage Ranking.
- Limitations:
 - SQuAD only for testing/evaluation purposes.
 - Use data, such that possible to use the BLEU evaluation method.
- Future Work:
 - Try other IR algorithms for getting most important passage.
 - Different models for open/closed generative.
 - Different evaluation metrics.

Q&A