Biomedical Question Answering

WILPS Master Project

Sushil Awale
Sara Mínguez Monedero
Pablo Robles de Zulueta

September 29, 2022

- Introduction
- System Overview
- Data Extraction
- Query Formulation
- Retrieval Model
- Question Answering Model
- User Interface
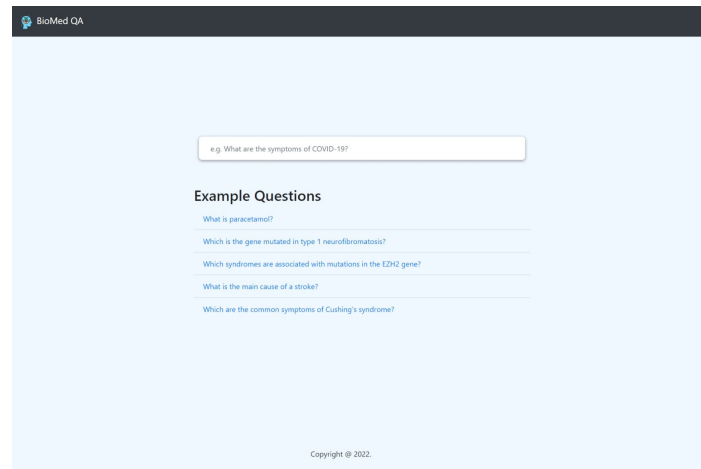- Demo
- Evaluation
- Discussion
- Conclusion

What it is?

IR and NLP system able to respond human's NLP biomedical queries.

How it can help nowadays society?

Yield biomedical quality information and source.

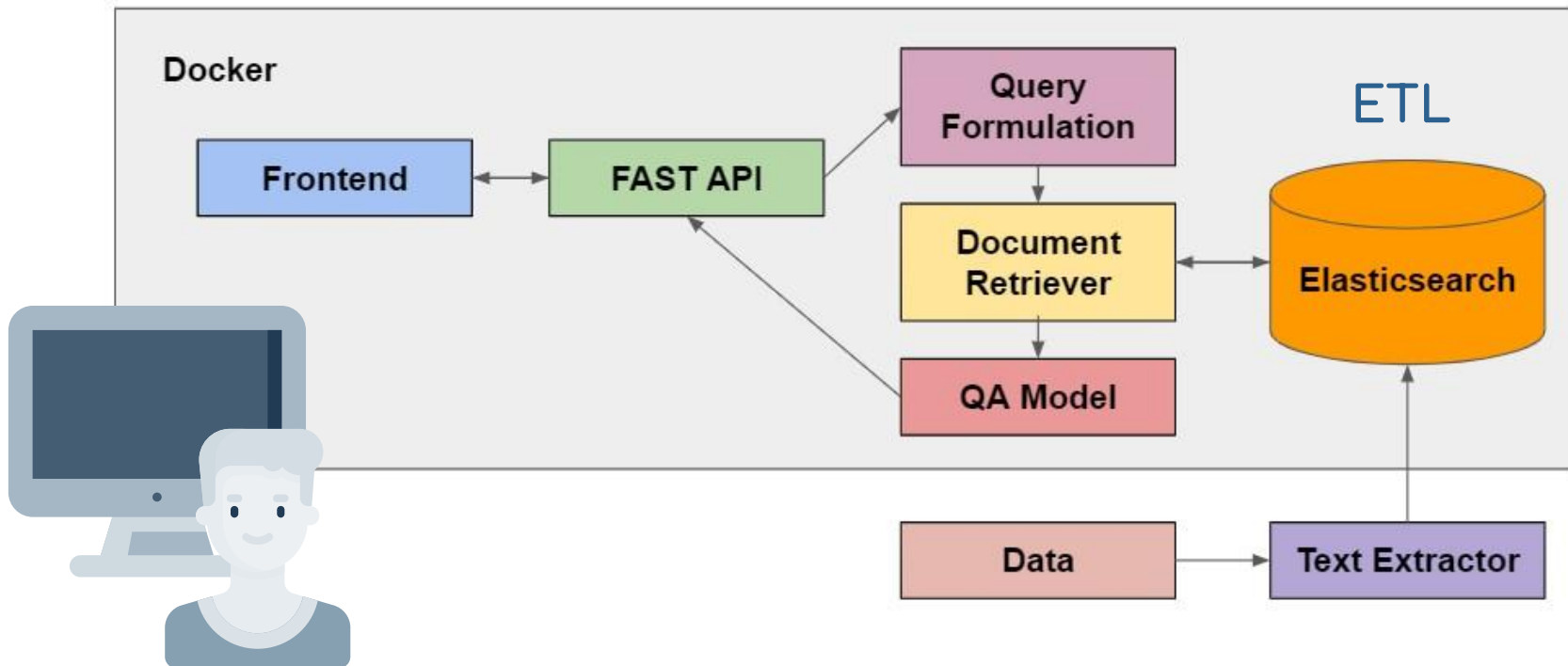Who is the target audience?

Medical field professionals & researchers.
General public.

Fig 1: Landing page of the QA system

Fig 2: BioMedQA System architecture diagram

- ■ **Used XML version of PubMed Open Access Non-Commercial Data Dump**
  - ○ About 173,000 research articles
  - ○ Extracted paragraphs and metadata
- ■ **BIOS**
  - ○ Data source to handle term definitions
  - ○ 248.345 definitions
- ■ **Totalling about 6.9 Million documents**

- **For PubMed research articles**
  - Remove citations using square brackets
  - Remove HTML tags
- **For BIOS definitions**
  - Filter out non-English definitions
  - Merge all related-concepts to one

    microglia, microglia cells and microglia cell

  - Re-format in the form

    term: definition text

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

01. PubMed.gov

National Center for Biotechnology Information's (NCBI) PubMed Central.

Open Access (OA) subset → **text** + **metadata**.

XML version of PubMed OA Non-Commercial Data Dump

About **173,000** research articles

02. BIOS

Biomedical Informatics Ontology System.

Data source to handle **term definitions**.

e.g., what are osteocytes?

7 million biomedical concepts matched to around **250 thousand definitions**.

Data Extraction - Formatting

0.1 RESEARCH ARTICLES

XML tags → allow easy paragraph extraction.

Metadata → title, author names, publication date and publication journal.

Remove citations ([ ]) and HTML tags in passages

02. CONCEPTS & DEFINITIONS

Filter out non-english terms.

Merge all related concepts to one
e.g., microglia, microglia cell, microglia cells

Re-format to store:
term : definition text

248.345 definitions

*Stored in ElasticSearch → about **6.9 Million** documents*

|		
Original question	*What's the most common clinical manifestation of Cytomegalovirus (CMV) infection AIDS patients?*	*I don't feel my hands, what can it be?*
Sentence cleaning	*what is the most common clinical manifestation of cytomegalovirus (cmv) infection in aids patients*	*i do not feel my hands, what can it be*
Medical entities	*'clinical', 'manifestation', 'cytomegalovirus (cmv) infection', 'patients', 'aids'*	*hands*
Noun chunks	*'clinical', 'manifestation', 'cytomegalovirus (cmv) infection', 'patients', 'aids'*	*hands*
Entity expansion	*most common clinical manifestation', 'aids patients', 'cytomegalovirus (cmv) infection'*	*hands*
Dependency parsing	*what, is*	*not, be, do, i, feel*
Final keywords	*common clinical manifestation cytomegalovirus cmv infection aids patients*	*not feel hands*

■ AIM → Extract relevant keywords.

■ Custom set of stop words → Preserve details

■ SpaCy (*en_core_web_sm*) + ScispaCy (*en_core_sci_lg*)

■ RegEx, SequenceMatcher and Contractions.

1. Sentence Preparation	2. SpaCy parsing	3. Medical entities extraction	4. Noun chunk extraction
5. Medical entity expansion	6. Dependency parsing	7. StopWords removal	8. Sentence arrangement

Original sentence	*What's the most common clinical manifestation of Cytomegalovirus (CMV) infection AIDS patients?*	*I don't feel my hands, what can it be?*
Sentence cleaning	*what is the most common clinical manifestation of cytomegalovirus (cmv) infection in aids patients*	*i do not feel my hands, what can it be*
Medical entities	*'clinical', 'manifestation', 'cytomegalovirus (cmv) infection', 'patients', 'aids'*	*hands*
Noun chunks	*'clinical', 'manifestation', 'cytomegalovirus (cmv) infection', 'patients', 'aids'*	*hands*
Entity expansion	*most common clinical manifestation', 'aids patients', 'cytomegalovirus (cmv) infection'*	*hands*
Dependency parsing	*what, is*	*not, be, do, i, feel*
Final sentence	*common clinical manifestation cytomegalovirus cmv infection aids patients*	*not feel hands*

Two step retrieval process:

- BM25 with Fuzzy Search
- Re-rank with semantic ranking
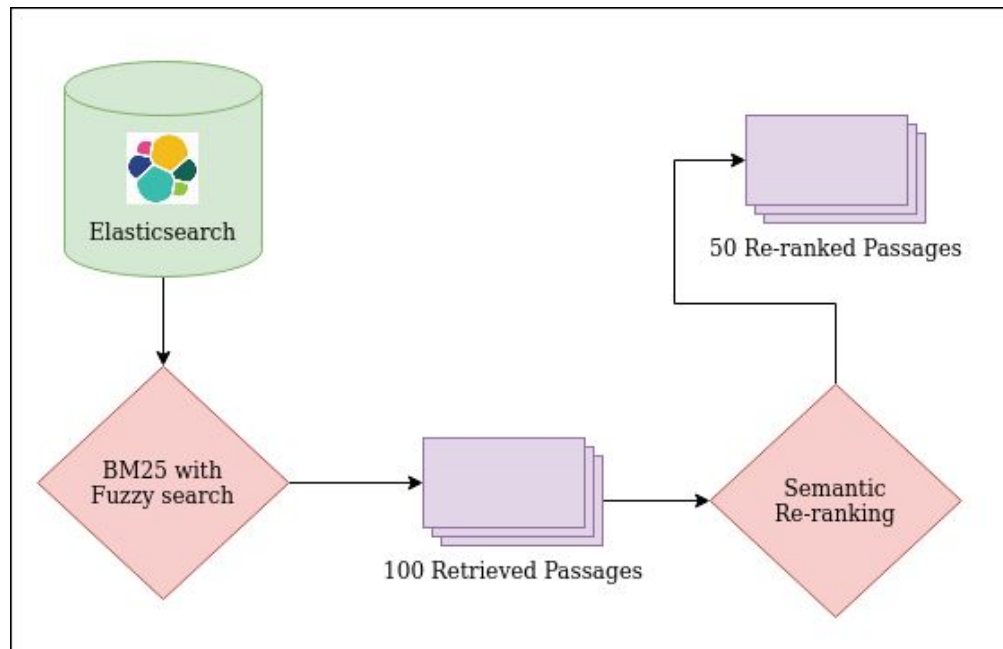
Fig 3: Passage retrieval process

- **BM25 with Fuzzy Search**
  - Search for keywords in titles and body
  - Use custom analyzer the handles lower case and English stop words
  - Allow for spelling errors up to 2 characters
  - Retrieve top 100 passages

- **Semantic Re-ranking**
  - Convert passages into embeddings using MiniLM [1]
  - MiniLM trained on MSMACRO dataset[1]
  - Re-rank the passages using Cosine similarity scores
  - Pass top 50 to QA Model

[1] https://github.com/microsoft/MSMARCO-Passage-Ranking
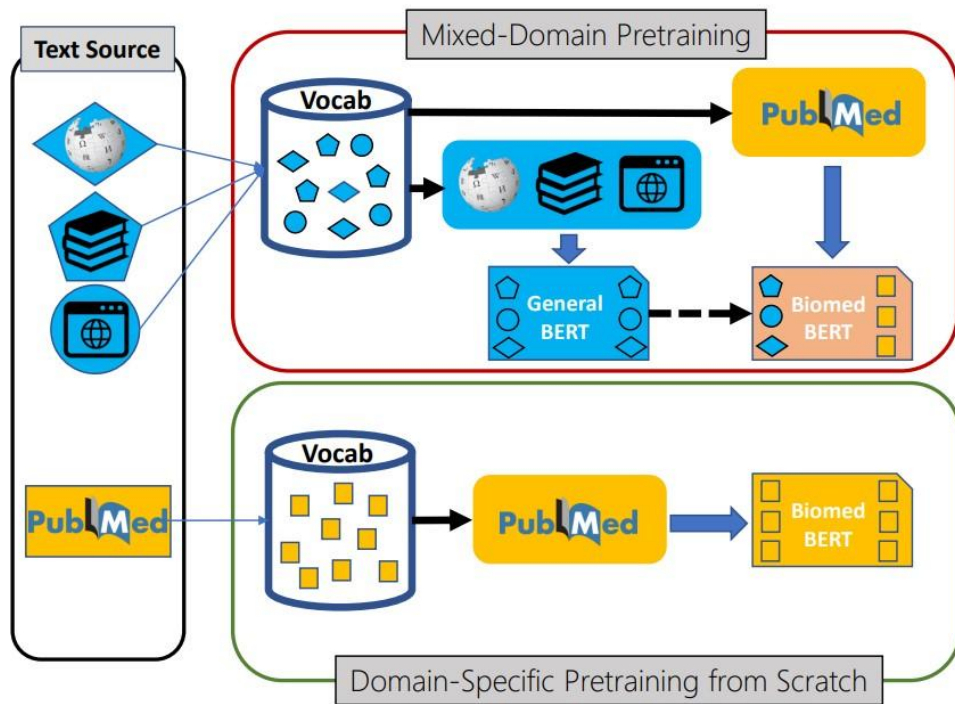
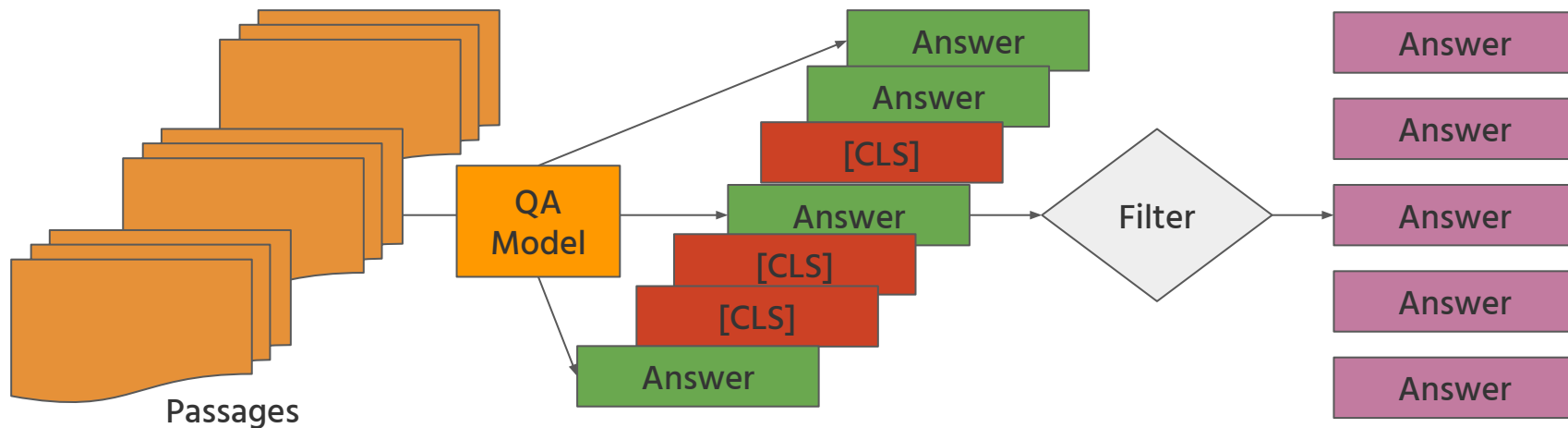*Fig 4: BERT Question Answering Model trained on PubMed [2]*

Use Fine tuned **PubMedBERT [2]** model on SQuAD v.2[2] dataset to build a neural QA model.

Advantages:

- Contains **vocabulary from PubMed** since it is pretrained from scratch.

- SQuAD v.2 dataset allows the model to **not find an answer** when the **context is not relevant.**
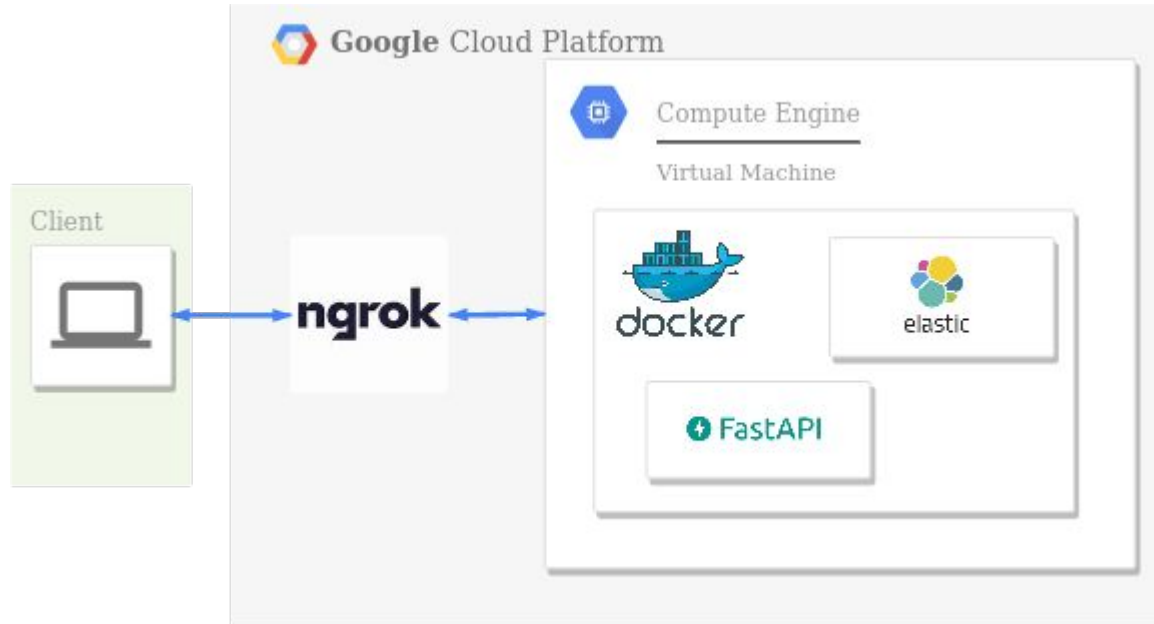
[2] https://rajpurkar.github.io/SQuAD-explorer/

14

Fig 5: Filtering out passages with no answers

Code: https://github.com/awalesushil/biomed-qa

Fig 6: BioMedQA System deployment architecture diagram

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Developed a webapp using **HTML** and **Bootstrap**.

- Accessible publicly using **NGROK** via a **web browser**.

- 5 **example questions** obtained from BioAsq[3] dataset shown on Landing Page

[3] http://participants-area.bioasq.org/datasets/
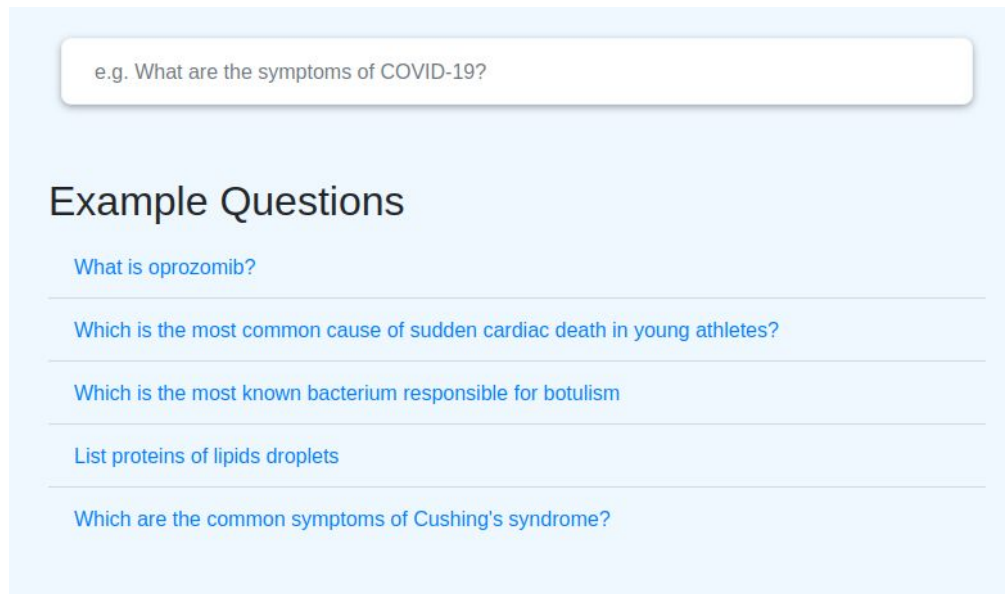
Fig 7: Screenshot of user interface, Landing page

Currently live at

https://bit.ly/biomedqa

■ For evaluation we design an evaluation system

    ○ **3 experts** in the field evaluate the model. **Label the answers** as *relevant* or *not-relevant* for 18 selected questions.

■ Questions compose of **5 different types**: *factoid, list, yes/no, how and why.*

■ Report the **precision@k** for **k = 1 to 5**

Sudden cardiac death athletes: a systematic review

Marcelo Ferreira, Paulo Roberto Santos-Silva, Luiz Carlos de Abreu, Vitor E Adriano Meneghini, Andrés R Pérez Riera, Tatiana Dias de Carvalho, Luiz Ca Filho, Celso Ferreira, 2010 - **Sports Medicine, Arthroscopy, Rehabilitation**

Context

For the vast majority of deaths caused by cardiovascular disease in athletes heart, i.e., diseases at birth. Among them the most commons are: hypertroph ventricular mass (10%). The remaining percentage is caused by other diseas

Read more

Answer *hypertrophic cardiomyopathy*

○ Not-relevant ● Relevant

Fig 8: Screenshot of user interface, Evaluation page

Table 2: Overall mean precision over the different types of question

Precision@rank	Mean Precision
P@1	0.5
P@2	0.5
P@3	0.42
P@4	0.46
P@5	0.58

Table 4: Overall mean precision over all factoid questions

Precision@rank	Mean Precision
P@1	0.7
P@2	0.7
P@3	0.699
P@4	0.7
P@5	0.7

NO MATCHING OF CONTEXT AND ANSWER

■ In some answers, although the answer is correct, the context provided casts doubt on this answer or refutes it.

BioMed QA

what is the treatment of acute myocarditis?

A systematic review of intravenous gamma globulin for therapy of acute myocarditis

Joan L Robinson, Lisa Hartling, Ellen Crumley, Ben Vandermeer, Terry P Klassen, 2005 - **BMC Cardiovascular Disorders**

Context

In conclusion, the value of IVGG in patients with acute myocarditis is obscured by the poor quality of evidence. We were not able to identify a subgroup of patients who appear to be more likely to respond to IVGG. A large RCT is required to evaluate the efficacy of IVGG for acute myocarditis with emphasis on the etiology of the myocarditis. Until there are RCTs demonstrating benefit, use of IVGG for acute myocarditis should not be part of routine practice. Moreover, there is a great need for further studies of the pathophysiology of acute myocarditis, which would allow for a better understanding of the etiology and the natural history of the disease. This might allow for improved diagnostic criteria, which would make it much easier to design studies of treatment options. This may also assist in identifying sub-groups of patients where IVGG or other therapies have a greater potential to confer clinical benefit.

Read less

Answer *ivgg*

22

- *Our QA model provides good answers for **medical technical questions** as well as **general knowledge medical questions**.*

- *It has a **good performance** for **factoid questions** and for other types when the **answer is explicitly provided**.*

- *QA model **fails to provide a suitable answer** when the tasks are **more NLP intensive** and require an increased level of language, semantics and syntactic reasoning.*

- ***Relevant tool** for medical field students or experts that could help to find relevant answers and articles **reducing the amount of time searching**.*

- Code: https://github.com/awalesushil/biomed-qa/

- Demo: https://bit.ly/biomedqa

- Report: https://github.com/awalesushil/biomed-qa/blob/main/report.pdf

Thank you!

Any Questions?

[1] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 11 2019.

[2] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," ACM Transactions on Computing for Healthcare (HEALTH), vol. 3, no. 1, pp. 1–23, 2021.