THE UNIVERSITY OF ALABAMA AT BIRMINGHAM

Cancer ASK

Authors:-

Mirza Tanzim Sami[1], Nilesh Kumar[2], Trupesh Patel[1]

Team name:

DeepBioComp

[1]Department of Computer Science, UAB, [2]Department of Biology, UAB

Presentation:-
Mirza Tanzim Sami,
Nilesh Kumar,
Trupesh Patel

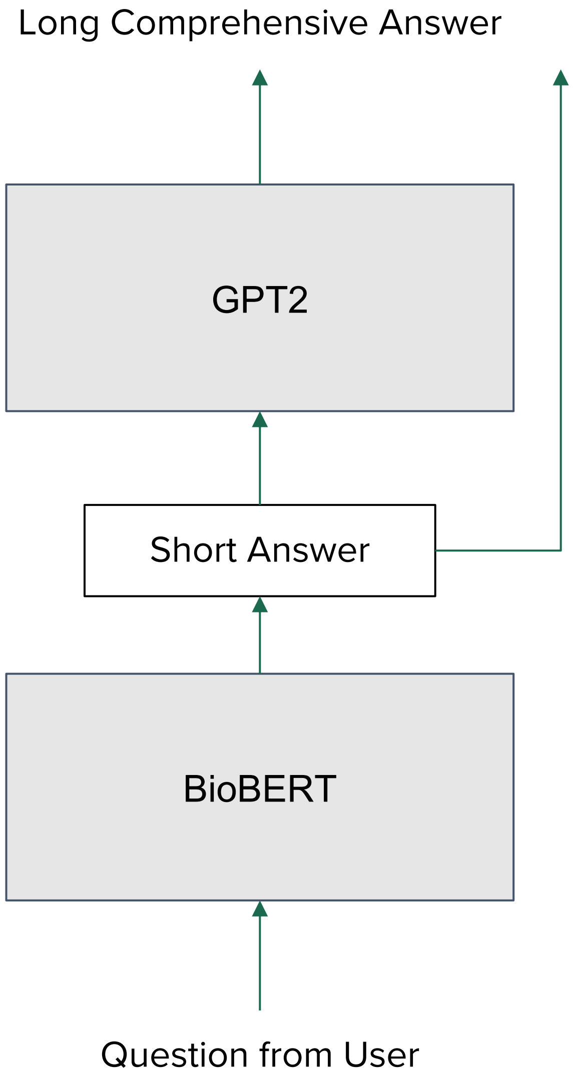
Instructor:-
Dr. John Osborne

Introduction

- Cancer Ask 🦀 is the **only** composite deep learning model for Cancer based question-answering task.
- It leverages two state of the art NLP based deep learning model
  - BioBERT
  - GPT2
- Cancer Ask 🦀 brings powerful question-answering capabilities, producing answers which are semantically and contextually meaningful.
- Answers can be short and sweet or verbose and comprehensive.

Motivation and Approach

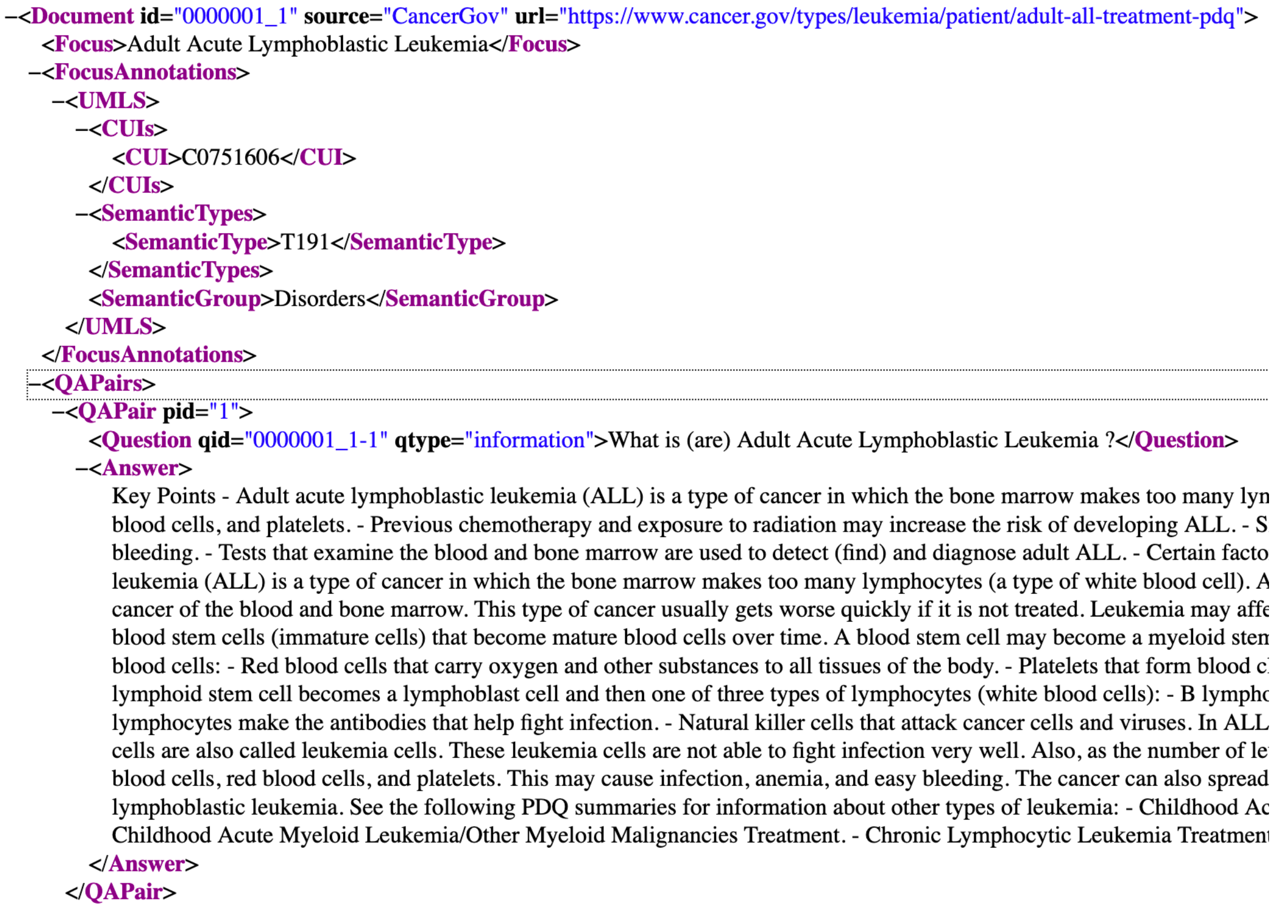
- Why?
- Fine-tuned on cancer dataset provided by MedQuAD dataset[1]
- Give short answers
- Produce context rich, comprehensive answers



Ref[1]: https://arxiv.org/abs/1901.08079

Dataset

- Cancer dataset, published as MedQuAD[1]
- Created for medical question-answering
- 12 NIH websites (e.g. cancer.gov, niddk.nih.gov, GARD, MedlinePlus Health Topics)
- Focused on ~100 types of cancers
- In total 729 questions
- 12 different type of questions



```
–<Document id="0000001_1" source="CancerGov" url="https://www.cancer.gov/types/leukemia/patient/adult-all-treatment-pdq">
   <Focus>Adult Acute Lymphoblastic Leukemia</Focus>
  –<FocusAnnotations>
    –<UMLS>
      –<CUIs>
          <CUI>C0751606</CUI>
        </CUIs>
      –<SemanticTypes>
          <SemanticType>T191</SemanticType>
        </SemanticTypes>
        <SemanticGroup>Disorders</SemanticGroup>
      </UMLS>
   </FocusAnnotations>
  –<QAPairs>
    –<QAPair pid="1">
        <Question qid="0000001_1-1" qtype="information">What is (are) Adult Acute Lymphoblastic Leukemia ?</Question>
      –<Answer>
          Key Points - Adult acute lymphoblastic leukemia (ALL) is a type of cancer in which the bone marrow makes too many lyn
          blood cells, and platelets. - Previous chemotherapy and exposure to radiation may increase the risk of developing ALL. - S
          bleeding. - Tests that examine the blood and bone marrow are used to detect (find) and diagnose adult ALL. - Certain facto
          leukemia (ALL) is a type of cancer in which the bone marrow makes too many lymphocytes (a type of white blood cell). A
          cancer of the blood and bone marrow. This type of cancer usually gets worse quickly if it is not treated. Leukemia may affe
          blood stem cells (immature cells) that become mature blood cells over time. A blood stem cell may become a myeloid sten
          blood cells: - Red blood cells that carry oxygen and other substances to all tissues of the body. - Platelets that form blood c
          lymphoid stem cell becomes a lymphoblast cell and then one of three types of lymphocytes (white blood cells): - B lympho
          lymphocytes make the antibodies that help fight infection. - Natural killer cells that attack cancer cells and viruses. In ALL
          cells are also called leukemia cells. These leukemia cells are not able to fight infection very well. Also, as the number of le
          blood cells, red blood cells, and platelets. This may cause infection, anemia, and easy bleeding. The cancer can also spread
          lymphoblastic leukemia. See the following PDQ summaries for information about other types of leukemia: - Childhood Ac
          Childhood Acute Myeloid Leukemia/Other Myeloid Malignancies Treatment. - Chronic Lymphocytic Leukemia Treatmen
        </Answer>
      </QAPair>
```

Dataset

Table 1. List of the question type and their count

Question type	Count
Information	112
Treatment	95
Susceptibility	88
Research	86
Symptoms	82
Exams and tests	82
Outlook	82
Stages	77
Prevention	12
Causes	7
Inheritance	5
Genetic changes	1

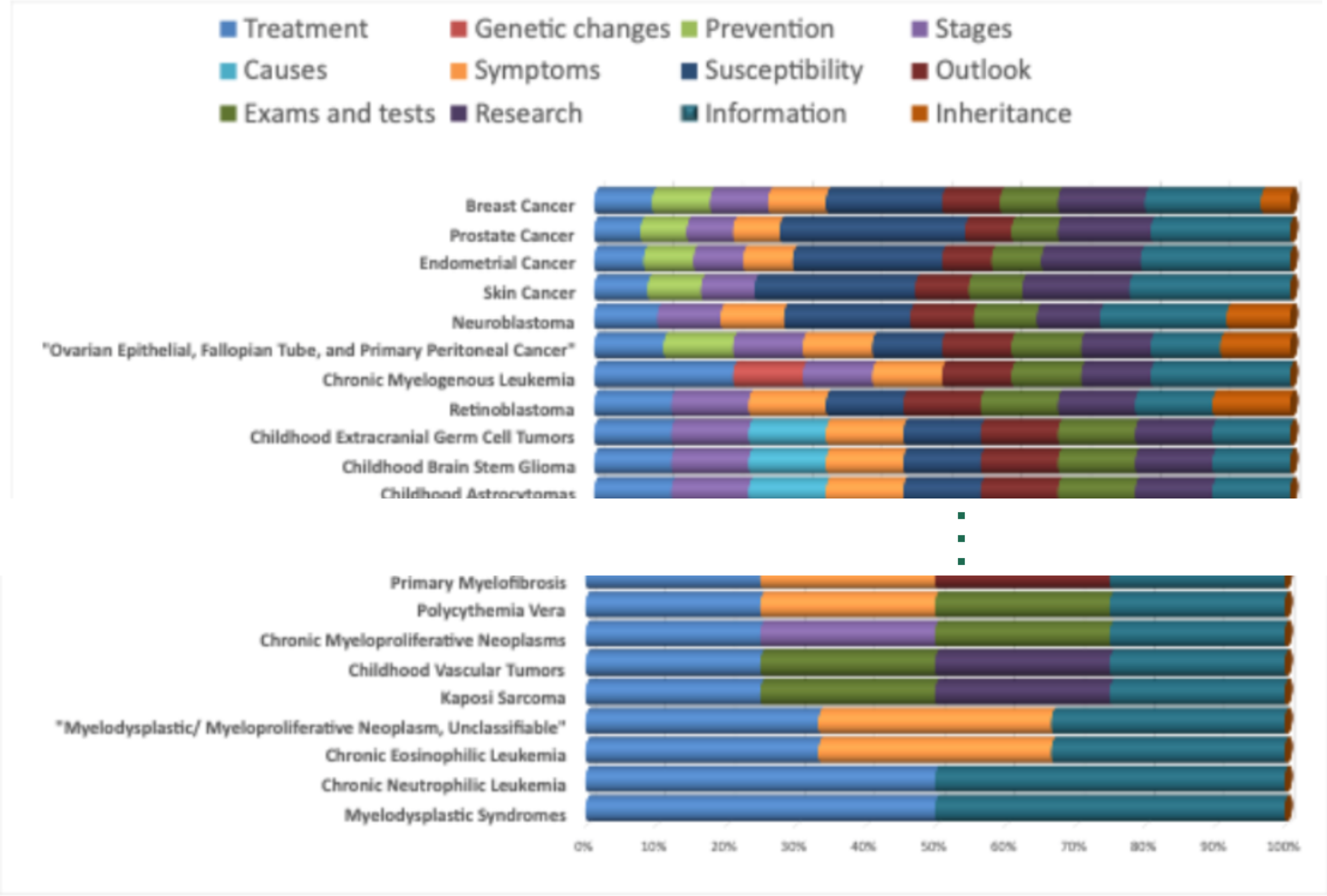


Fig. 1. List of the Cancer type and question type and their count

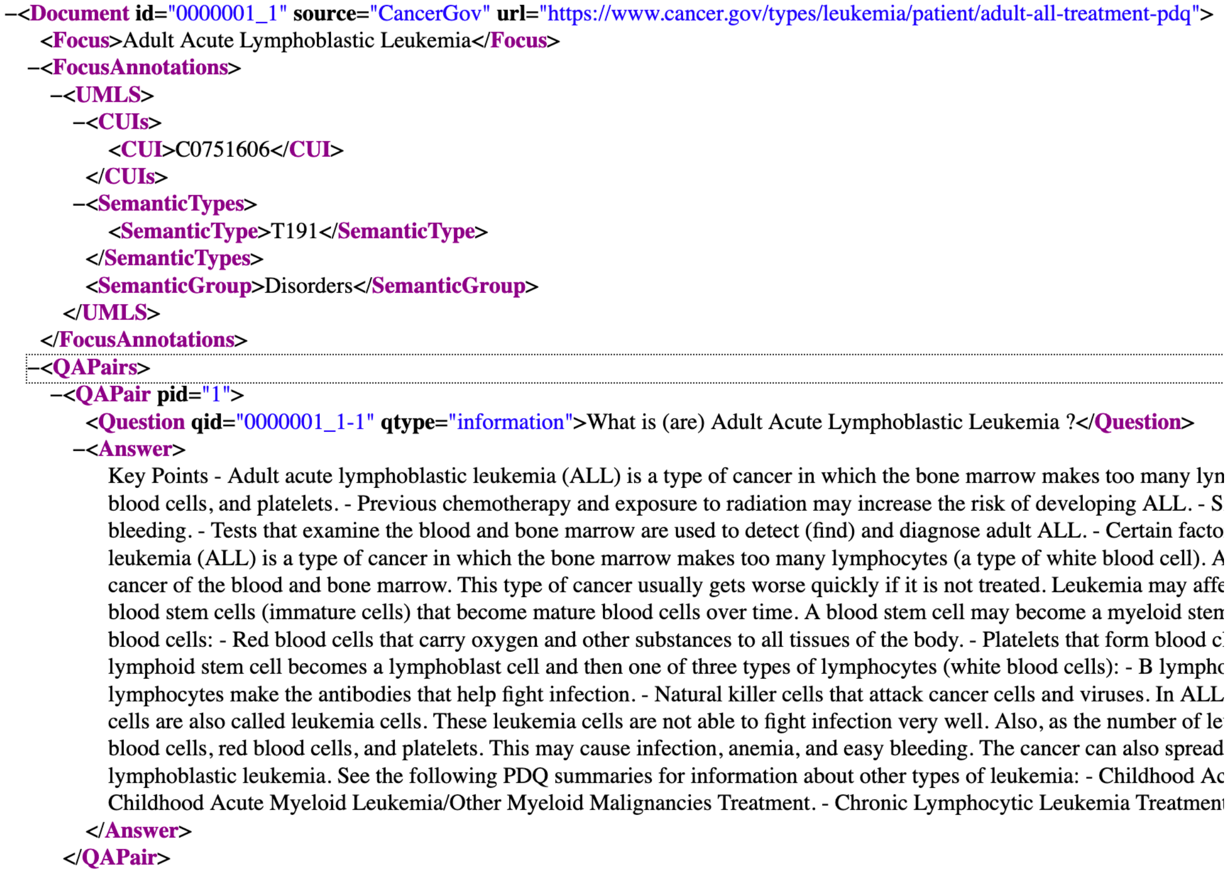
Dataset Challenges

- BioBERT model require tags:
  - Context
  - Question
  - Answer
  - Start_answer
- Missing tags in raw data:
  - Context
  - Start_answer

- Data modification steps:
  - Answer (XML tag) → Context (JSON key)
  - Context (JSON key) to short and correct Answer (JSON key)
  - Calculate Start_answer (JSON Key) index from Context (JSON key)

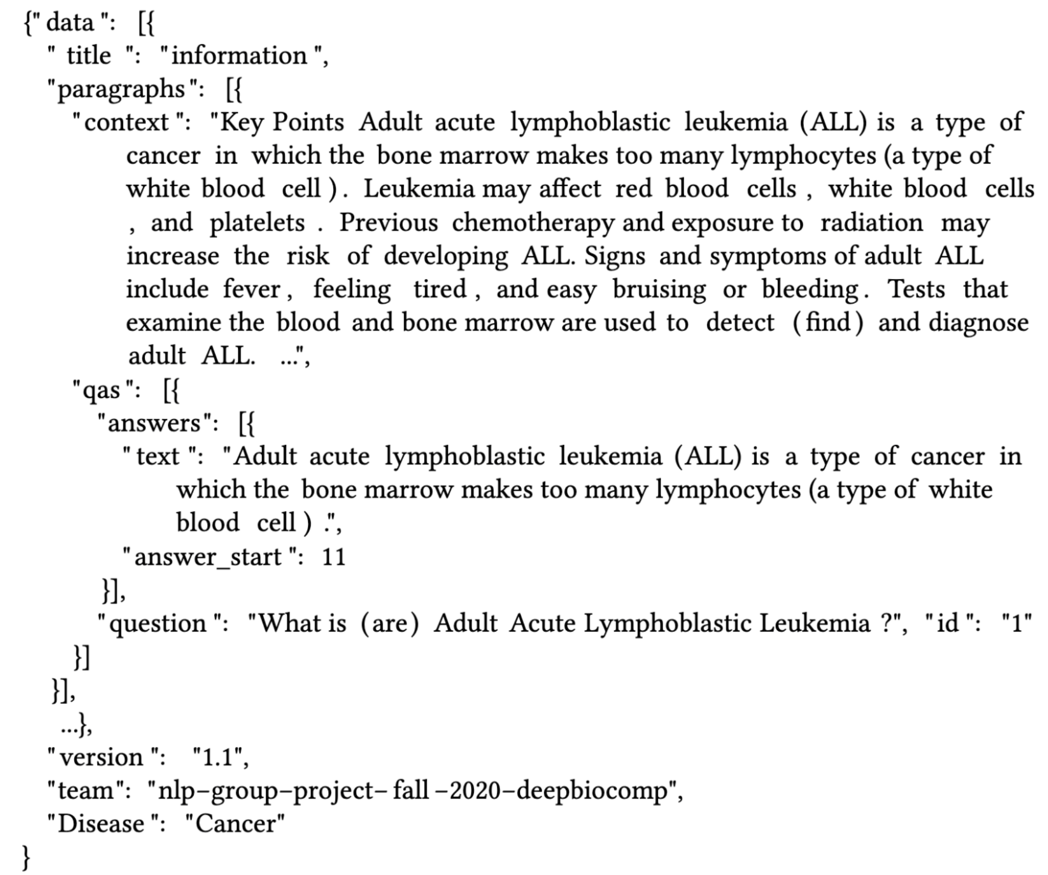
Dataset challenge

Before



```
–<Document id="0000001_1" source="CancerGov" url="https://www.cancer.gov/types/leukemia/patient/adult-all-treatment-pdq">
   <Focus>Adult Acute Lymphoblastic Leukemia</Focus>
  –<FocusAnnotations>
    –<UMLS>
      –<CUIs>
          <CUI>C0751606</CUI>
        </CUIs>
      –<SemanticTypes>
          <SemanticType>T191</SemanticType>
        </SemanticTypes>
        <SemanticGroup>Disorders</SemanticGroup>
      </UMLS>
   </FocusAnnotations>
  –<QAPairs>
    –<QAPair pid="1">
        <Question qid="0000001_1-1" qtype="information">What is (are) Adult Acute Lymphoblastic Leukemia ?</Question>
      –<Answer>
          Key Points - Adult acute lymphoblastic leukemia (ALL) is a type of cancer in which the bone marrow makes too many lyn
          blood cells, and platelets. - Previous chemotherapy and exposure to radiation may increase the risk of developing ALL. - S
          bleeding. - Tests that examine the blood and bone marrow are used to detect (find) and diagnose adult ALL. - Certain facto
          leukemia (ALL) is a type of cancer in which the bone marrow makes too many lymphocytes (a type of white blood cell). A
          cancer of the blood and bone marrow. This type of cancer usually gets worse quickly if it is not treated. Leukemia may affe
          blood stem cells (immature cells) that become mature blood cells over time. A blood stem cell may become a myeloid stem
          blood cells: - Red blood cells that carry oxygen and other substances to all tissues of the body. - Platelets that form blood c
          lymphoid stem cell becomes a lymphoblast cell and then one of three types of lymphocytes (white blood cells): - B lympho
          lymphocytes make the antibodies that help fight infection. - Natural killer cells that attack cancer cells and viruses. In ALL
          cells are also called leukemia cells. These leukemia cells are not able to fight infection very well. Also, as the number of le
          blood cells, red blood cells, and platelets. This may cause infection, anemia, and easy bleeding. The cancer can also spread
          lymphoblastic leukemia. See the following PDQ summaries for information about other types of leukemia: - Childhood Ac
          Childhood Acute Myeloid Leukemia/Other Myeloid Malignancies Treatment. - Chronic Lymphocytic Leukemia Treatmen
        </Answer>
      </QAPair>
```

After



```
{" data ":  [{
  " title  ":  " information ",
  " paragraphs ":  [{
    " context ":  " Key Points  Adult  acute  lymphoblastic  leukemia  (ALL) is  a  type  of
        cancer  in  which the  bone marrow makes too many lymphocytes (a type of
        white blood  cell ) .  Leukemia may affect  red blood  cells ,  white blood  cells
        ,  and  platelets  .  Previous  chemotherapy and exposure to  radiation  may
        increase  the  risk  of  developing  ALL. Signs  and symptoms of adult  ALL
        include  fever ,  feeling  tired ,  and easy  bruising  or  bleeding .  Tests  that
        examine the  blood  and bone marrow are  used  to  detect  ( find )  and diagnose
        adult  ALL.   ...",
    " qas ":  [{
      " answers ":  [{
        " text ":  " Adult  acute  lymphoblastic  leukemia  (ALL) is  a  type  of  cancer  in
            which the  bone marrow makes too many lymphocytes (a type of  white
            blood  cell ) .",
        " answer_start ":  11
      }],
      " question ":  " What is  ( are )  Adult  Acute Lymphoblastic Leukemia ?",  " id ":  " 1 "
    }]
  }],
  ...},
  " version ":   " 1.1 ",
  " team":  " nlp−group−project− fall −2020−deepbiocomp",
  " Disease ":  " Cancer "
}
```

BioBERT

- A variant of the BERT model.
- BioBERT comes fine-tuned on the PubMed dataset.
- Fine-tuned again on MedQuAD dataset to make more precise toward cancer specific QA.
- The script is modified from the ‘run_squad.py’ script from Huggingface’s Transformer[2].

[2] Ref: https://github.com/huggingface/transformers/blob/master/examples/question-answering/run_squad.py

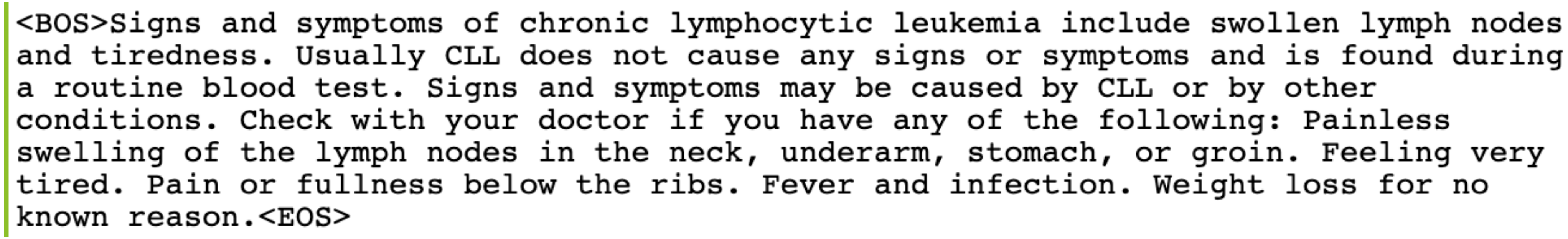
BioBERT

- Fine – tuned parameter:
  - "model_name_or_path": **'dmis-lab/biobert-base-cased-v1.1-squad**' ,
  - "learning_rate": 5e-5 ,
  - "num_train_epochs": 20 ,
  - "max_seq_length": 384 ,
  - "doc_stride": 128 ,
  - "per_gpu_eval_batch_size": 12 ,
  - "per_gpu_train_batch_size": 12 ,
  - "save_steps": 5000,
  - "do_lower_case": True,

Ref[2]: https://github.com/huggingface/transformers/blob/master/examples/question-answering/run_squad.py

GPT2

- Unsupervised model
- Uses only Context from the dataset
- Every sentence
  - <BOS> (beginning of sentence) tag
  - <EOS> (end of sentence) tag
- The script is modified from the 'run_language_modelling.py' script from Huggingface's Transformer git repo.



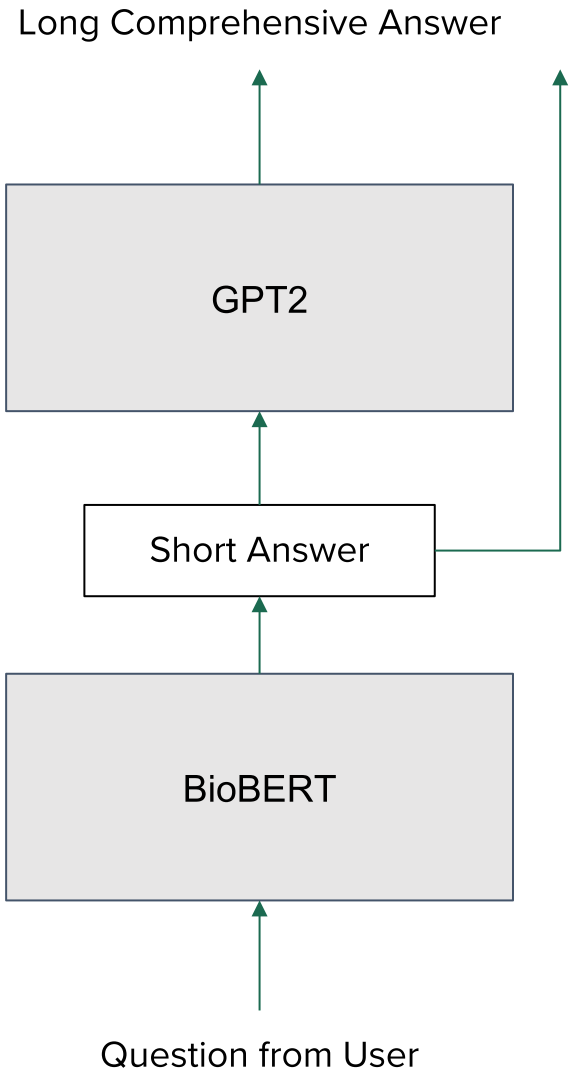
```
<BOS>Signs and symptoms of chronic lymphocytic leukemia include swollen lymph nodes
and tiredness. Usually CLL does not cause any signs or symptoms and is found during
a routine blood test. Signs and symptoms may be caused by CLL or by other
conditions. Check with your doctor if you have any of the following: Painless
swelling of the lymph nodes in the neck, underarm, stomach, or groin. Feeling very
tired. Pain or fullness below the ribs. Fever and infection. Weight loss for no
known reason.<EOS>
```

GPT2

- Fine – tuned parameter:
  - "model_name_or_path": ‘**gpt2**’ ,
  - "learning_rate": 5e-5 ,
  - "num_train_epochs": 20 ,
  - "per_gpu_eval_batch_size": 1 ,
  - "per_gpu_train_batch_size": 1 ,
  - "save_steps": 100,
  - "eval_steps": 100
  - "line_by_line": true

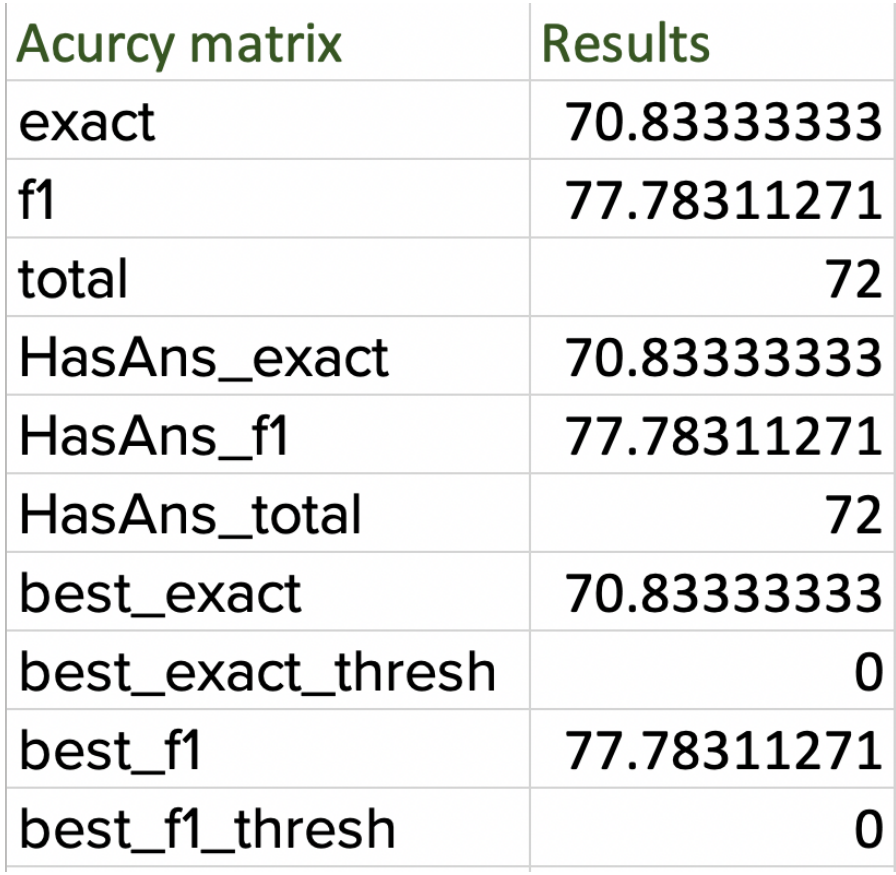
Composite Model

- The BioBERT model returns an answer for the given query, the output is used as the input for the GPT2 model.
- The text-generation code is based on the 'run_generation.py' script from the Huggingface's Transformer git repository.
- The script has been completely modified and rewritten as a python function in a Jupyter notebook. It has also been modified to accept our dataset which contains the '<BOS>' and '<EOS>' special tokens.
- Thus the model provides a detailed answer to cancer related queries submitted by user.



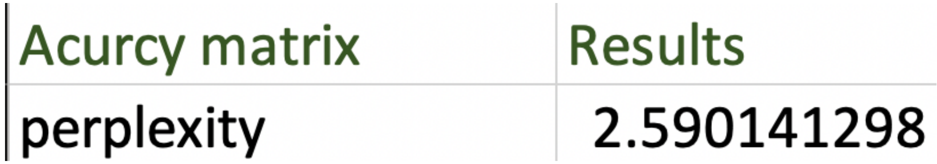
Accuracy Metrics

- BioBERT Tuning:



Acurcy matrix	Results
exact	70.83333333
f1	77.78311271
total	72
HasAns_exact	70.83333333
HasAns_f1	77.78311271
HasAns_total	72
best_exact	70.83333333
best_exact_thresh	0
best_f1	77.78311271
best_f1_thresh	0

- GPT2 Tuning:



Acurcy matrix	Results
perplexity	2.590141298

Qualitative Results

- Example 1
  - **Question:** "What are the stages of Vulvar Cancer ?"
  - **Short Answer:** stage i stage ii stage iii stage iv after vulvar cancer has been diagnosed, tests are done to find out if cancer cells have spread within the vulva or to other parts of the body.
  - **Long Answer:** *stage i stage ii stage iii stage iv after vulvar cancer has been diagnosed, tests are done to find out if cancer cells have spread within the vulva or to other parts of the body. The following stages are used for vulvar cancer: Stage I Stage II Stage III Stage IV Stage V After vulvar cancer has been diagnosed, tests are done to find out if cancer cells have spread within the vulva or to other parts of the body. The process used to find out if cancer has spread within the vulva or to other parts of the body is called staging. The information gathered from the staging process determines the stage of the disease. It is important to know the stage in order to plan treatment. Vulvar cancer in men may affect the risk of developing rectal cancer. Having rectal cancer during surgery can increase the risk of rectal cancer and is called a risk factor. Having rectal cancer in a first degree relative (parent, brother, or sister) increases the risk even more.*

Qualitative Results

- Example 2
  - **Question:** "What are the symptoms of the Ovarian Germ Cell Tumors ?"
  - **Short Answer:** signs of ovarian germ cell tumor are swelling of the abdomen or vaginal bleeding after menopause.
  - **Long Answer:** *signs of ovarian germ cell tumor are swelling of the abdomen or vaginal bleeding after menopause. These and other signs and symptoms may be caused by ovarian germ cell tumor or by other conditions. Check with your doctor if you have any of the following: Swelling of the abdomen or vaginal bleeding after menopause. Trouble starting the flow of urine. Weight loss for no known reason. Trouble emptying the bladder completely after menopause. Pain or feeling of fullness below the ribs on the left side. Fever for no known reason. Easy bruising or bleeding. Pain or feeling of fullness below the ribs on the right side.*

Short-comings & Future work

- For the ease of fine-tune on available resources, only initial part of the context was used.
- Further, the transformer-based models like BioBERT are inefficient to handle long sequences due to expensive self-attention operations.
- To address this limitation, we will fine-tune the Longformer model and compare results against our model.
- Since the attention mechanism in the Longformer scales linearly with sequence length, it may reflect weaknesses of our model.
- Additionally, the chosen dataset (transformer-based) is too small and we were unable to procure the dataset from BioASK. When available we will finetune over this big data set to add more diversity to the model.

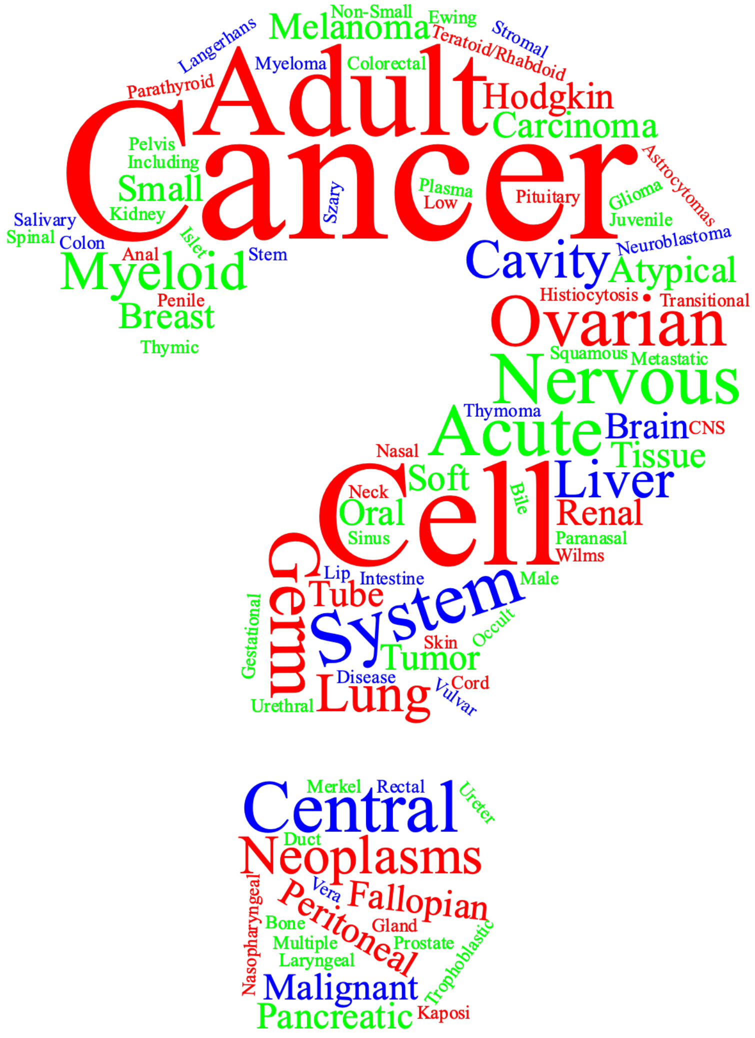
Conclusion

- The answers or responses produced by our fine-tuned model support the relevance of the question in logic with cancer QA.
- Our model also highlights the possibility of combining text generative and long-former models for further model improvement.
- Our findings also show that relying on a restricted set of reliable answer sources can bring a plentiful improvement in domain-specific QA.

References

- https://github.com/abachaa/MedQuAD
- https://github.com/huggingface/transformers
- https://huggingface.co/dmis-lab/biobert-base-cased-v1.1-squad
- https://huggingface.co/transformers/model_doc/gpt2.html

Q & A


Non-Small
Melanoma
Ewing
Langerhans
Stromal
Teratoid/Rhabdoid
Myeloma
Colorectal
Parathyroid
Adult
Hodgkin
Carcinoma
Cancer
Pelvis
Including
Small
Kidney
Szary
Plasma
Low
Pituitary
Astrocytomas
Glioma
Juvenile
Salivary
Spinal
Colon
Islet
Anal
Stem
Neuroblastoma
Cavity
Atypical
Myeloid
Penile
Histiocytosis
Transitional
Breast
Ovarian
Squamous
Metastatic
Thymic
Nervous
Thymoma
Brain
CNS
Acute
Nasal
Tissue
Soft
Liver
Neck
Bile
Cell
Oral
Renal
Sinus
Paranasal
Wilms
Lip
Intestine
Male
Germ
Tube
System
Gestational
Skin
Occult
Tumor
Disease
Cord
Lung
Vulvar
Urethral
Merkel
Rectal
Ureter
Central
Duct
Neoplasms
Vera
Fallopian
Peritoneal
Nasopharyngeal
Bone
Gland
Multiple
Prostate
Laryngeal
Trophoblastic
Malignant
Pancreatic
Kaposi

UAB THE UNIVERSITY OF ALABAMA AT BIRMINGHAM

Thank you!