# **Abstract**

With the advent of new sophisticated experiment and analysis technology in medicine, the data generation in medicine has accelerated by several folds since the last decades. A big part of all gathered knowledge is the collection of text documents such as research articles. Keeping abreast of biomedical research developments, several efficient natural language processing (NLP) text mining models have also developed. Though, often it is difficult to apply those text mining models directly to domain-specific biomedical corpora. In this project, we combined two different types of NLP models so that the combined model can give a comprehensive answer to cancer-specific questions. At first, we fine-tuned BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) model for short question types over cancer-related question answers and context texts, we called *Cancer Ask*. Further, we fine-tuned the Generative Pre-trained Transformer 2 (GPT2), models over cancer-related text, we named it as *GPT2 Cancer*. The fundamental objective of this project is to use answers from *Cancer* *Ask*, feed it to *GPT2 Cancer* and provide a comprehensive answer to cancer related queries.

# **1. Materials and methods**

The following sections are categorized into three main sections, first a brief explanation on the data processing and implementation methods used to fine-tune the BioBERT model for regular question answering. Second, the data processing and fine-tuning of the GPT2 model for text-generation which is the basis for creation of long comprehensive answers. Finally, the report discusses the implementation methods for a composite model that runs the fine-tuned BioBERT for Question-Answering and uses the output of the BioBERT model as the input prompt for the GPT2 text generator for a more verbose, comprehensive answer.

## **1.1 Fine-tuning BioBERT**

The biomedical domain texts contain a vast number of domain-specific proper nouns (e.g. BRCA1, Leukemia) which are understood mostly by biomedical researchers. In this context the BioBERT is already fine-tuned on PubMed abstracts (PubMed) and PubMed Central full-text articles(PMC). Further, we fine-tuned again using Cancer QA dataset, the MedQuAD. The total Cancer and question type is summarized in table 1.

**Table 1. List of the Cancer type and question type.**

|  |  |
| --- | --- |
| Cancer type | Question type |
| Ovarian Epithelial, Fallopian Tube, and Primary Peritoneal Cancer | 10 |
| Breast Cancer | 10 |
| Anal Cancer | 9 |
| Adult Central Nervous System Tumors | 9 |
| Childhood Astrocytomas | 9 |
| Childhood Brain Stem Glioma | 9 |
| Endometrial Cancer | 9 |
| Childhood Extracranial Germ Cell Tumors | 9 |
| Retinoblastoma | 9 |
| Neuroblastoma | 9 |
| Prostate Cancer | 9 |
| Adult Acute Myeloid Leukemia | 8 |
| Chronic Myelogenous Leukemia | 8 |
| Hairy Cell Leukemia | 8 |
| Childhood Acute Myeloid Leukemia and Other Myeloid Malignancies | 8 |
| Adult Soft Tissue Sarcoma | 8 |
| Childhood Soft Tissue Sarcoma | 8 |
| Adult Hodgkin Lymphoma | 8 |
| Adult Non-Hodgkin Lymphoma | 8 |
| Childhood Hodgkin Lymphoma | 8 |
| Childhood Central Nervous System Atypical Teratoid/Rhabdoid Tumor | 8 |
| Childhood Central Nervous System Germ Cell Tumors | 8 |
| Childhood Craniopharyngioma | 8 |
| Childhood Ependymoma | 8 |
| Adult Primary Liver Cancer | 8 |
| Bile Duct Cancer (Cholangiocarcinoma) | 8 |
| Childhood Liver Cancer | 8 |
| Osteosarcoma and Malignant Fibrous Histiocytoma of Bone | 8 |
| Gastrointestinal Carcinoid Tumors | 8 |
| Uterine Sarcoma | 8 |
| Extragonadal Germ Cell Tumors | 8 |
| Intraocular (Uveal) Melanoma | 8 |
| Gallbladder Cancer | 8 |
| Gestational Trophoblastic Disease | 8 |
| Langerhans Cell Histiocytosis | 8 |
| Hypopharyngeal Cancer | 8 |
| Laryngeal Cancer | 8 |
| Lip and Oral Cavity Cancer | 8 |
| Nasopharyngeal Cancer | 8 |
| Oropharyngeal Cancer | 8 |
| Paranasal Sinus and Nasal Cavity Cancer | 8 |
| Salivary Gland Cancer | 8 |
| Pancreatic Cancer | 8 |
| Wilms Tumor and Other Childhood Kidney Tumors | 8 |
| Male Breast Cancer | 8 |
| Skin Cancer | 8 |
| Melanoma | 8 |
| Merkel Cell Carcinoma | 8 |
| Non-Small Cell Lung Cancer | 8 |
| Parathyroid Cancer | 8 |
| Penile Cancer | 8 |
| Pituitary Tumors | 8 |
| Colon Cancer | 8 |
| Rectal Cancer | 8 |
| Adult Acute Lymphoblastic Leukemia | 7 |
| Chronic Lymphocytic Leukemia | 7 |
| Childhood Acute Lymphoblastic Leukemia | 7 |
| Childhood Rhabdomyosarcoma | 7 |
| Mycosis Fungoides and the Szary Syndrome | 7 |
| Childhood Non-Hodgkin Lymphoma | 7 |
| Childhood Brain and Spinal Cord Tumors | 7 |
| Childhood Central Nervous System Embryonal Tumors | 7 |
| Ewing Sarcoma | 7 |
| Ovarian Germ Cell Tumors | 7 |
| Ovarian Low Malignant Potential Tumors | 7 |
| Metastatic Squamous Neck Cancer with Occult Primary | 7 |
| Pancreatic Neuroendocrine Tumors (Islet Cell Tumors) | 7 |
| Transitional Cell Cancer of the Renal Pelvis and Ureter | 7 |
| Plasma Cell Neoplasms (Including Multiple Myeloma) | 7 |
| Small Cell Lung Cancer | 7 |
| Small Intestine Cancer | 7 |
| Thymoma and Thymic Carcinoma | 7 |
| Urethral Cancer | 7 |
| Vulvar Cancer | 7 |
| Gastrointestinal Stromal Tumors | 6 |
| AIDS-Related Lymphoma | 6 |
| Primary CNS Lymphoma | 6 |
| Testicular Cancer | 6 |
| Chronic Myelomonocytic Leukemia | 5 |
| Kaposi Sarcoma | 4 |
| Childhood Vascular Tumors | 4 |
| Liver (Hepatocellular) Cancer | 4 |
| Chronic Myeloproliferative Neoplasms | 4 |
| Polycythemia Vera | 4 |
| Primary Myelofibrosis | 4 |
| Essential Thrombocythemia | 4 |
| Myelodysplastic/ Myeloproliferative Neoplasms | 4 |
| Juvenile Myelomonocytic Leukemia | 4 |
| Atypical Chronic Myelogenous Leukemia | 4 |
| Ovarian, Fallopian Tube, and Primary Peritoneal Cancer | 4 |
| Oral Cavity and Oropharyngeal Cancer | 4 |
| Renal Cell Cancer | 4 |
| Colorectal Cancer | 4 |
| Chronic Eosinophilic Leukemia | 3 |
| Myelodysplastic/ Myeloproliferative Neoplasm, Unclassifiable | 3 |
| Lung Cancer | 3 |
| Myelodysplastic Syndromes | 2 |
| Chronic Neutrophilic Leukemia | 2 |

**Table 2. 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 |

## **1.2 Fine-tuning GPT2**

This section will briefly explain the data used to pretraint the GPT2 model and how the script that was implemented to use the GPT2 model for text generation. The GPT2 model was chosen for this task because of the robustness of the model and its ability to generate long sentences while maintaining relatively good semantic sense. However, the language model is much too general and requires fine-tuning to work effectively and generative texts pertinent to cancer queries. Thus, the model was fine-tuned on the same PubMed dataset that was used to fine-tune the aforementioned BioBERT model.

However, unlike the BioBERT model which is used for question-answering, the data must be processed differently for the GPT2 model. In the case of the question-answering models, the dataset usually has three major components, the ‘question’, ‘context’ and the ‘answer’. All of which are important for training a question-answering model, but fine-tuning a language model (GPT2) does not require all three components. Concretely, the GPT2 model was trained only using the ‘context’ of the dataset. Furthermore, two special tokens were added, the ‘<BOS>’ signifying the beginning of a sentence and a ‘<EOS>’ token, signifying the end of a sentence. The code for fine-tuning the GPT2 model is based on an older (not currently available) script from the huggingface transformer git repo called ‘run\_language\_modelling.py’. The script as modified and reimplemented as a jupyter notebook. Furthermore, it was also modified to let GPT2 accept the special tokens as mentioned above. Once the model was fine-tuned with our desired dataset, it was saved locally to be used for can related text-generation.

## **1.3 Composite Model for Comprehensive Question Answering**

The final section will discuss the composite model that stacks the fine-tuned GPT2 based text-generation model on top of the fine-tuned BioBERT model for question-answering. The question-answering code is based on the ‘run\_squad.py’ script from the **huggingface transformer** git repository. The fine-tuned model takes as input the question provided by the user and tries to give an answer that is correct and contextually relevant to the question asked. Once, 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 transformer** git repository. The script has been heavily 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. The text-generation model returns two suitable answers which are generated based on the prompt provided from the question-answering model. The resulting final answer is not only verbose and comprehensive but semantically and contextually relevant to the question asked. Thus, providing a much better experience to the user submitting the queries to the model.

# **2. Training & Results**

This section of the report will discuss the results obtained from the two models used in this project, the metrics used to measure the accuracy of the model and the qualitative measure of the generated answers.

## **2.1 Training and Accuracy Metrics**

First, the BioBERT model was fine-tuned on our dataset. The model was trained for 20 epoch running on 4 GPU cores. The model reduced the loss value with each epoch during training. Then, the following accuracy metrics scores were achieved during evaluation on test data.

'exact': 70.83333333333333,

'f1': 77.78311271345326,

'total': 72,

'HasAns\_exact': 70.83333333333333,

'HasAns\_f1': 77.78311271345326,

'HasAns\_total': 72,

'best\_exact': 70.83333333333333,

'best\_exact\_thresh': 0.0,

'best\_f1': 77.78311271345326,

'best\_f1\_thresh': 0.0

The accuracy metrics were monitored carefully to ensure that the model generalizes well and not overfit to the training data.

Next, the GPT2 model was fine-tuned on the data processed for it. This model was tuned for 19 epochs and the loss value decreased with each epoch. Once the model was tuned, the accuracy of the model was evaluated on the text data. The primary metric used was the perplexity score of the model which is as follows, 'perplexity': 2.5901412982272793. Again, special attention is given to ensure that the model remains general and not overfit to the data.

## **2.2 Qualitative Results**

Given the nature of the project, numeric accuracy scores may not be an intuitive indicator of the performance of the model. Thus, a few examples are given below that shows the question asked to the model, the answer given by the BioBERT model and the verbose, comprehensive answer given by the GPT2 model stacked on top.

### **Set 1:**

Question: "What are the symptoms of…rian Germ Cell Tumors ?"

Answer: signs of ovarian germ cell tumor are swelling of the abdomen or vaginal bleeding after menopause.

Comprehensive Answer1: 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: Weakness or feeling tired. Weight loss with little or no effect on menstrual periods. Vaginal bleeding after menopause. Sometimes pain or swelling in a women's vagina. Fever or night sweats.

Comprehensive Answer2: 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.

Comprehensive Answer3: signs of ovarian germ cell tumor are swelling of the abdomen or vaginal bleeding after menopause. Other conditions can increase the risk of hairy cell leukemia. Check with your doctor if you have any of the following: Pain or swelling in the abdomen. Weakness or feeling tired. Weight loss for no known reason. A menstrual period that does not go away.

Comprehensive Answer4: signs of ovarian germ cell tumor are swelling of the abdomen or vaginal bleeding after menopause. These and other signs may be caused by ovarian germ cell tumor or by other conditions. Check with your doctor if you have any of the following: Pain or swelling in the abdomen. A lump in the abdomen, vagina, or rectum. Weight loss for no known reason. Pain or a feeling of fullness below the ribs on the left side. Weight loss with no known reason. A dark urinelike color that does not go away.

### **Set 2:**

Question: = "What are the stages of Vulvar Cancer ?"

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.

Comprehensive Answer1: 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. Cancer that has spread within the vulva to other parts of the body is called seminoma. A biopsy is done to diagnose seminoma. A biopsy is done during surgery to remove the tumor. Sometimes, a small incision (cut) is made in the vulva and a laparoscope (a thin, lighted tube) is inserted into it. The doctor will usually remove the tumor and a small amount of tissue, after the doctor has removed all the cancer. The tissue sample is checked under a microscope for signs of cancer. A pathologist views the tissue under a microscope to look for cancer cells. If cancer cells are not found, it may not be necessary to remove more tissue. There are three ways that cancer spreads in the body. Cancer can spread through tissue, the lymph system, and the blood: Tissue. The cancer spreads from where it began by growing into nearby areas. Lymph system. The cancer spreads from where it began by getting into the lymph system. The cancer travels through the lymph vessels to other parts of the body. Blood. The cancer spreads from where it began by getting into the blood. The cancer travels through the blood vessels to other parts of the body. Cancer may spread from where it began to other parts of the body through organs and tissues. When cancer spreads to another part of the body, it is called metastasis. Cancer cells break away from where they began (the primary tumor) and travel through the lymph system or blood.

Comprehensive Answer2: 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.

## **2.3 Future Plans**

(Nilesh, I’ll leave this to you, Please talk about Longformers here. Thanks)

# **3. Conclusions**

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 effectiveness 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**