MedGen: A Python Natural Language Processing Toolkit for Medical Text Processing

Rohith Chowadry Singamaneni   
 *department of computer science*

*California State EastBay* Hayward, CA, USA  
rohithsingamanenirc@gmail.com

Murali Krishna Kancheti  
*department of computer science*

*California State EastBay*Hayward, CA, USA  
kanchetimuralikrishna24@gmail.com

*Abstract*— This study introduces MedGen, an extensive natural language processing (NLP) toolkit specifically crafted for the processing of medical text. Designed with the needs of biomedical researchers and healthcare professionals in mind, MedGen offers an accessible, all-encompassing solution that requires minimal programming expertise. Its features include:

Generative Functions: MedGen introduces four advanced generative functions, namely question answering, text summarization, text simplification, and machine translation.

Basic NLP Functions: MedGen seamlessly integrates 12 fundamental NLP functions, covering tasks such as word tokenization and sentence segmentation.

Query and Search Capabilities: MedGen provides user-friendly query and search functions tailored for text corpora exploration.

Bottom of Form

Keywords—generative functions, tokenization, sentence segmentation, nlp ,health-care,summarization,gpt

Materials and Methods

We conducted a meticulous fine-tuning process for 32 domain-specific language models, subjecting them to thorough evaluations across 24 established benchmarks. To ensure practical relevance, we engaged clinicians in manual reviews. Furthermore, we enhanced our toolkit by introducing intuitive query and search functions. Additionally, we standardized and seamlessly integrated functions from third-party libraries, augmenting the overall capabilities of our toolkit.

Results:

| **Toolkit** | **\*Question Answering** | **Text Summarization** | **Text Simplification** | **Machine Translation** | **Basic NLP Functions** | **Query Search** |
| --- | --- | --- | --- | --- | --- | --- |
| MIMIC-Extract20 |  |  |  |  |  |  |
| scispaCy21 |  |  |  |  |  |  |
| MedspaCy22 |  |  |  |  |  |  |
| Transformers-sklearn23 |  |  |  |  |  |  |
| Stanza Biomed24 |  |  |  |  |  |  |
| MedGen |  |  |  |  |  |  |

The consistently enhanced performance of the fine-tuned models was particularly evident in text generation tasks. Notably, it led to a significant improvement of 9.75 and 3.37 in ROUGE-2 scores for text summarization and text simplification tasks, respectively. In the answer generation task, manual reviews indicated that the generated answers achieved impressive average scores of 4.95 (out of 5) in Readability, 4.43 in Relevancy, 3.9 in Accuracy, and 3.31 in Completeness.

***Conclusions:***

*Presented in this study is the creation and assessment of MedGen, an NLP toolkit specifically crafted for the straightforward processing of medical text. MedGen provides a comprehensive, user-friendly solution that encompasses four generative functions, twelve fundamental NLP functions, and intuitive search and query capabilities.*

# Introduction

The realm of medical texts poses considerable domain-specific challenges, characterized by issues like ambiguities, frequent abbreviations, the prevalence of negations, and complexities in segmentation. The manual curation of these texts stands as a time-consuming and labor-intensive endeavor. In response to these challenges, there has been a growing reliance on natural language processing (NLP) algorithms for automating text processing.

Recent years have witnessed a significant transition from shallow embeddings like BioWordVec and BioSentVec to sophisticated architectures such as Bidirectional Encoder Representations from Transformers (BERT), exemplified by domain-specific pre-trained language models like BioBERT, ClinicalBERT, and PubMedBERT. This shift has substantially elevated the efficacy of NLP tasks within biomedical and clinical domains, encompassing areas such as text classification, named entity recognition, text segmentation, language translation, and text generation.

Despite the effectiveness of these advanced techniques, a discernible gap persists between their sophistication and practical utilization by downstream users, specifically biomedical researchers, and healthcare professionals. The technical intricacies pose substantial challenges for direct application, particularly for individuals lacking a background in computational methods or basic programming skills. As a result, there is an increasing need for toolkits that are user-friendly and accessible, aiming to streamline the complexities associated with medical text processing.

Multiple toolkits are available for text processing in the biomedical domain. Table 1 summarizes representative tools.

**Table 1** presents a comprehensive comparison with existing toolkits, with the ⋆ symbol denoting tasks evaluated through human assessment. Basic NLP Functions, encompassing abbreviation extraction, sentence tokenization, word tokenization, negation detection, hyponym detection, UMLS concept extraction, named entity recognition, document clustering, POS tagging, entity linking, text summarization (extractive methods), and multi-choice QA, are considered. It is important to note that while not every toolkit incorporates all 12 basic NLP functions, MedGen distinguishes itself by including each of these functionalities.

This table offers an insightful overview, highlighting the distinctions and commonalities among various toolkits, with a specific emphasis on the inclusion of essential NLP functions where MedGen stands out for its comprehensive coverage.

MIMIC-Extract functions as a comprehensive pipeline designed for data extraction, preprocessing, and representation from the MIMIC-III dataset. scispaCy, on the other hand, serves as a tool that customizes spaCy's models to effectively process scientific and biomedical text. MedspaCy, also built on the spaCy framework, offers a combination of rule-based and machine learning-based methods tailored for processing medical text. Transformers-sklearn23 provides a toolkit facilitating the seamless integration of pre-trained Transformer-based models into the scikit-learn framework. Stanza Biomed stands out as an advanced tool addressing statistical, neural, and rule-based challenges in computational linguistics. It offers a straightforward interface for NLP tasks, showcasing nearly state-of-the-art performance using neural networks. Despite the strengths of these existing toolkits, their emphasis on different perspectives and the absence of generation capabilities in any of them create a notable gap.

In response to existing gaps, we introduce MedGen, an extensive NLP toolkit specifically tailored for medical text processing. MedGen marks a pioneering step by incorporating four advanced generative functions: question answering, text summarization, text simplification, and machine translation. Additionally, MedGen encompasses 12 fundamental NLP functions, including word tokenization and sentence segmentation, along with intuitive query and search capabilities.25 Furthermore, we conducted the fine-tuning of 32 domain-specific language models, subjecting them to rigorous evaluations across 24 established benchmarks, and solicited manual reviews from two healthcare professionals. MedGen is designed to empower a broad spectrum of users, ranging from novices to seasoned professionals, enabling them to effortlessly address their NLP tasks, even with limited technical expertise in handling textual data. We firmly believe that MedGen not only democratizes access to cutting-edge methods but also accelerates their seamless integration into healthcare practices.

**MATERIALS AND METHODS**

MedGen stands as a specialized and all-encompassing NLP toolkit meticulously designed for the processing of medical text. This toolkit comprises three integral modules:

**Generative Functions**: Serving as the cornerstone of MedGen, this module encompasses four generative tasks: question answering, text summarization, text simplification, and machine translation. These functions cater to diverse application scenarios within the healthcare domain.

**Basic NLP Functions:** This module encompasses 12 fundamental NLP functions, forming a crucial component of MedGen's capabilities.

**Query and Search Capabilities:** MedGen provides user-friendly query and search functions tailored for efficient exploration of text corpora.

The holistic architecture of MedGen is visually represented in Figure 1, illustrating the interconnectedness and integration of its key components.

A screenshot of a computer program

Description automatically generated

Figure 1 illustrates the comprehensive architecture of MedGen. The ⚙️ symbol signifies the presence of fine-tuned models dedicated to specific tasks, showcasing the toolkit's specialization. Additionally, tasks marked with ⭐️ indicate instances where thorough evaluations have been conducted, underscoring the commitment to rigorous assessment and refinement within MedGen.Bottom of Form

***Generative Functions***

MedGen leverages pre-trained language models, providing users with a suite of capabilities such as question answering, text summarization, text simplification, and machine translation. Users also have the flexibility to tap into any publicly available language models. Furthermore, we offer fine-tuned models tailored for specific generative tasks, all of which are publicly accessible for users to reference and integrate into their workflows. In the subsequent sections, we will delve into each of these potent generative functions individually.

## **Question Answering**

In the realm of healthcare, question answering emerges as especially pivotal. Its integration into healthcare systems plays pivotal roles, including pre-consultation and remote consultation, adeptly managing the surge in patient volume, and easing the burden on the healthcare infrastructure. Furthermore, dedicated question answering systems possess the potential to make significant contributions to patient education and medical training. Within MedGen, we have integrated the question answering function, encompassing two distinct sub-tasks: multiple-choice question answering and answer generation.

### **Multiple-Choice Question Answering**

We equip users with a biomedical multiple-choice question answering function, enabling them to input both the question text and options to ascertain the most likely answer. While the primary methodology involves a classification approach, we have innovatively employed generative models in the role of encoders, thus placing this task within this category. Our comprehensive comparative analysis extended to five pre-trained language models in the biomedical and clinical domain, namely BioBERT, ClinicalBERT, SapBERT, GatorTron-base, and PubMedBERT. These models underwent fine-tuning and evaluation using the Head-QA and MedMCQA datasets. Head-QA encompasses questions across six topics: medicine, nursing, psychology, chemistry, pharmacology, and biology, all sourced from professional position exams within the Spanish healthcare system. Additionally, MedMCQA represents a larger dataset covering 2,400 healthcare topics and 21 medical subjects. Owing to the absence of labels in the MedMCQA test set, we utilized the validation set for evaluation purposes.

### **Answer Generation**

In addition to the multiple-choice question answering task, we also offer the capability of answer generation. For this particular task, we employed Baize-healthcare and OPT-MedQuAD, both pretrained on the MedQUAD dataset. MedQUAD comprises 47,457 question-answer pairs in the medical domain derived from 12 National Institutes of Health (NIH) websites. Evaluations were performed using the QA Test Collection from the TREC-2017 LiveQA medical task, containing 2,479 questions and their corresponding reference answers. Notably, recognizing the limitations of objective metrics in accurately assessing the quality of generated content, we conducted a manual validation in subsequent sections. Two healthcare professionals conducted a manual review of 50 randomly selected answers.

## **Text Summarization**

In the realm of healthcare, practitioners and researchers grapple with an exponential influx of information, spanning literature, Electronic Health Records (EHRs), and more. Text summarization emerges as a pivotal generative task, endeavoring to distill vital insights from the intricate fabric of texts and present them in a more condensed format.36 Automated text summarization proves instrumental for clinicians and researchers, enabling them to efficiently glean essential information while mitigating the challenges of information overload.

Within this module, we furnish an abstractive text summarization function. We conducted comparisons across general pretrained summarization models, including Pegasus, BigBird, BART, and PRIMERA, fine-tuned on general text summarization corpora. Additionally, we explored domain-specific models such as SciFive and BioBART, tailored to biomedical corpora like Pubmed and PMC. For evaluation purposes, we selected datasets such as PubMed, MIMIC-CXR,, and MEDQA-AnS.

The PubMed dataset encompasses 133,000 biomedical scientific publications from the PubMed database, with each input document representing a scientific article and the reference summarization being the associated abstract. MIMIC-CXR, on the other hand, comprises a de-identified dataset of chest radiographs in DICOM format, accompanied by free-text radiology reports. We utilized a subset from MIMIC-CXR for the MEDIQA 2021 Radiology report summarization shared task. Due to unavailability of the test set, we repurposed the validation set as the test set and additionally extracted 2000 instances from the training set to create a new validation set. MEDQA-AnS serves as a collection of consumer health questions and passages containing information pertinent to the questions, supporting both single-document and multiple-document summarization evaluation.

***Text Simplification***

Biomedical texts are often rife with intricate terminologies, posing a challenge for individuals lacking a clinical background. Within MedGen, the text simplification function aims to transform complex and technical biomedical texts into more understandable content. This enhancement facilitates improved comprehension and engagement for non-clinical individuals, including patients. By rendering the information more accessible, individuals can actively participate in clinical decisions with greater effectiveness.

Our evaluation involved several pre-trained models, namely BigBirdPegasus, BART, and BioBART, assessed on datasets such as eLife, PLOS, and MedLane. eLife and PLOS constitute shared task data released from the BioLaySumm 2023 Task 1, where the objective is to generate lay summarizations given longer inputs. While eLife and PLOS originated from the shared task, we encountered challenges in obtaining the ground truth of the original test set. To ensure a fair comparison, we conducted testing on the development dataset, reserving some examples from the original training set for validation.

MedLane, on the other hand, represents a comprehensive human-annotated dataset featuring professional-to-customer sentences selected from MIMIC-III. In the case of MedLane, we segregated 2,030 examples from the training set to form the validation set, employing the original test set for subsequent evaluation. Fine-tuning procedures were applied to selected pre-trained models, including Pegasus, BART, and BioBART.

Bottom of Form

## **Machine Translation**

Language barriers present formidable challenges for patients, hindering their access to timely information and effective communication with healthcare providers, ultimately resulting in diminished healthcare service quality. In addressing this issue, our machine translation function within MedGen endeavors to translate text from a source language to a target language in clinical scenarios. Leveraging pre-trained models, MedGen extends support to 17 languages.

To enhance the translation capabilities, we fine-tuned existing models such as MarianMT and multilingual T5 using the UFAL Medical Corpus. This corpus encompasses diverse medical text sources, including titles of medical Wikipedia articles, medical term-pairs, patents, and documents from the European Medicines Agency. During the preprocessing phase, we excluded general domain data from UFAL, such as parliamentary proceedings, and meticulously shuffled the medical-domain corpora. Subsequently, we split the data into two parts, allocating 85% for training and 15% for testing. For each language pair, we harnessed all available parallel data to optimize the breadth and accuracy of our machine translation function.

# ***Basic NLP Functions***

Within this module, we seamlessly incorporate numerous third-party libraries, offering comprehensive support for 12 distinct functions. These functions encompass abbreviation extraction, sentence tokenization, word tokenization, negation detection, hyponym detection, UMLS concept extraction, named entity recognition, document clustering, POS tagging, entity linking, text summarization (extractive method), and multi-choice QA. Further details can be found in Supplementary Appendix B. For evaluation purposes, we specifically chose to assess the performance of POS tagging and named entity recognition tasks.

# ***Query and Search Capabilities***

MedGen facilitates user-friendly query and search functions on text corpora through the following mechanisms:

*MySQL Support for MIMIC-III Database:* The data tables, specifically NOTEEVENTS.TSV, were meticulously indexed into a MySQL database. We offer user-friendly interfaces that empower users with basic statistical functions, allowing them to retrieve counts of patients, documents, and sentences.

*Query Functionality:* MedGen implements a variety of straightforward query functions. Users, for example, can retrieve a specified number of patient records or notes by utilizing their respective IDs.

*Search Capability:* Recognizing the pivotal role of effective search functionality within unstructured text, we integrated keyword search capabilities. These capabilities, supported by multiple libraries, ensure swift and targeted searches, enhancing the overall user experience.

# Results

## Question Answering and Multiple -Choice question answering

In our fine-tuning process, we leveraged five biomedical pre-trained models: BioBERT, ClinicalBERT, SapBERT, GatorTron-base, and PubMedBERT. The evaluation metric employed was the accuracy score, as depicted in Figure 2. The results highlight PubMedBERT's standout performance on HEAD-QA and MedMCQA (without context), achieving accuracy rates of 42.52% and 46.59%, respectively. On MedMCQA (with context), SapBERT, PubMedBERT, and GatorTron-base demonstrate comparable performance, with GatorTron-base emerging as the top performer, boasting an impressive accuracy of 64.93%.

A graph of the number of patients

Description automatically generated

**Figure 2.** Evaluation for multiple-choice question answering task.

## Answer Generation

The evaluation of the answer generation capabilities of two pre-trained models, Baize-healthcare and OPT-MedQuAD, was conducted using ROUGE scores53. Baize-healthcare outperformed OPT-MedQuAD across all R-1, R-2, and R-L scores, with respective scores of 21.11, 5.14, and 19.27. However, it's worth noting that the chosen metrics may not comprehensively assess the quality of healthcare-generated content. To address this limitation, manual reviews were performed by two healthcare professionals, evaluating content based on Readability, Relevancy, Accuracy, and Completeness. The detailed results of these reviews are provided in the manual validation section.Bottom of Form

## Text Summerization

The evaluation of text summarization encompassed both single-document and multi-document scenarios, with the assessment based on ROUGE scores. In Table 2, we compared the performance of five selected models across four chosen benchmarks for the single-document scenario. To ensure a fair comparison, results from BioBART and SciFive, which were fine-tuned on PubMed, were excluded. The observations indicate that BART consistently exhibits strong performance across three benchmarks. Notably, BART outperforms both SciFive and BioBART in terms of competitiveness. Furthermore, BioBART surpasses BART on only one of the benchmarks.

A table with numbers and letters

Description automatically generated

**Table 2.** Evaluation for single-document summarization. Some results are derived from other papers.

Additionally, our evaluation extends to multi-document summarization, utilizing the MEDQA-AnS dataset, as illustrated in Table 3. A diverse set of models, encompassing both traditional and deep learning approaches, were compared. TextRank was employed for extractive summarization, while abstractive summarization considered models such as BART, Pegasus, PRIMERA, and BioBART. Notably, BART displayed competitive performance, with BioBART showing slightly inferior results.

A table with numbers and letters

Description automatically generated

## Text Simplification

We conducted a comparison among BigBirdPegasus, BART, and BioBART, fine-tuning them specifically for text simplification tasks. The evaluation, depicted in Figure 3 (A), utilized ROUGE scores. Interestingly, both BART and BioBART exhibited superior performance to BigBirdPegasus across all three datasets. Notably, BioBART, which is pre-trained on biomedical corpora atop BART, demonstrated a marginally better performance on a single dataset.

图表, 条形图

描述已自动生成

**Figure 3.** (A) Evaluation for text simplification task using ROUGE scores. (B) Evaluation for text simplification task using FKGL score.

Moreover, we conducted an analysis of reading ability using the Flesch-Kincaid grade level (FKGL) score. The FKGL score serves as a measure of text complexity, indicating the difficulty of understanding a given text, as illustrated in Figure 3 (B). We compared the outputs generated by our models with the ground truth. For the eLife and PLOS datasets, the ground truth displayed FKGL scores of 12 and 15, respectively. Intriguingly, while the BioBART model demonstrated competitive performance in terms of ROUGE metrics, it failed to significantly reduce the difficulty of understanding, as evidenced by its FKGL score of 17 in both datasets. On the other hand, the BART model managed to slightly lower the FKGL score to 14 and 16 for eLife and PLOS, respectively. However, in the case of the MedLane dataset, all methods appeared to reach a similar level of complexity as the ground truth. This could be attributed to the dataset's shorter examples and potentially smaller vocabulary size, limiting the observed differences.

## Machine Translation

We fine-tuned MarianMT and mT5 on three language pairs: "en-es", "en-fr", "en-ro", utilizing MarianMT as the baseline for comparison. Evaluation was performed using BLEU score57, as depicted in Figure 4. Following fine-tuning, there was a significant improvement in BLEU scores, with the most substantial enhancement observed in the "en-fr" language pair. This improvement can be attributed to the larger amount of training data available for "en-fr" (2,812,305 samples). Furthermore, across all three language pairs, the mT5 model consistently outperformed the MarianMT model in terms of BLEU scores. Additionally, we conducted fine-tuning for mT5 on five more language pairs: "en-cs", "en-de", "en-hu", "en-pl", and "en-sv";

图表, 条形图

描述已自动生成

**Figure 4.** Evaluation for machine translation task.

# Improvement

Successful implementation of the MedGen API and OpenAI API integration. The integration combines the capabilities of MedGen, a comprehensive natural language processing (NLP) toolkit designed for medical text processing, with the advanced language model GPT-3.5 from OpenAI. The collaborative use of these APIs creates a powerful solution for addressing a wide range of medical text processing tasks.

**1. MedGen API Integration:**

The MedGen API serves as the backbone for medical text processing tasks. It provides a user-friendly interface for tasks such as text summarization, question answering, text simplification, and machine translation. The API is accessed through a chatbot platform, enabling users to interact seamlessly with MedGen's capabilities.

**2. OpenAI API Integration (GPT-3.5):**

The OpenAI API, specifically leveraging the GPT-3.5 language model, enhances the conversational and context-aware aspects of the system. GPT-3.5 is utilized for generating follow-up questions, providing detailed explanations, and assisting with additional context in user interactions.

**3. Chatbot Platform:**

A chatbot platform, chosen based on project requirements and preferences (e.g., Dialogflow, Microsoft Bot Framework, Rasa), serves as the intermediary between users and the integrated APIs. The platform processes user input, determines intent, and appropriately invokes the MedGen and OpenAI APIs.

**4. User Input and Interaction:**

Users interact with the system using natural language input related to their medical text processing needs. The chatbot processes user queries, understands the underlying intent, and seamlessly directs requests to either the MedGen API or the OpenAI API.

**5.Processing Flow:**

* **User Input Analysis:** The chatbot analyzes user input to discern the specific medical text processing task required.
* **MedGen API Invocation:** For tasks like text summarization, question answering, or text simplification, the chatbot invokes the relevant functions in the MedGen API.
* **OpenAI API Interaction (Optional):** Optionally, GPT-3.5 is engaged to provide additional context-aware responses, generating follow-up questions, or enhancing the conversational aspect.
* **User Response:** The system compiles the responses from both MedGen and GPT-3.5 to form a cohesive and informative reply, which is then presented to the user.

**Benefits and Significance**

The integration of the MedGen API and OpenAI API brings forth several key benefits:

1. **Enhanced Capabilities:** The collaborative use of MedGen and GPT-3.5 expands the range of tasks the system can handle, providing users with a more comprehensive solution for medical text processing.
2. **Context-Aware Interactions:** The inclusion of GPT-3.5 enables the system to better understand and respond to user input with context-awareness, creating a more natural and engaging interaction.
3. **User-Friendly Interface:** The chatbot platform acts as a user-friendly interface, allowing individuals, including those without advanced technical skills, to effortlessly interact with the integrated APIs.
4. **Scalability and Adaptability:** The modular architecture of the integration allows for scalability and adaptability, making it well-suited for future enhancements, updates, and additional features.

# Future Work

Antegration with Electronic Health Records (EHR) is a critical advancement that aims to streamline and enhance the accessibility of patient data within healthcare systems. By exploring compatibility with EHR systems, the integration seeks to directly process and extract information from electronic health records, offering several benefits for healthcare professionals and patients alike.

**EHR Compatibility:** The integration with EHR involves creating a seamless connection between the text processing system, such as MedGen, and the electronic health records stored in healthcare databases. This allows for the automatic extraction of relevant medical information, enabling more efficient and accurate responses to user queries. Healthcare providers can quickly access patient histories, diagnoses, and treatment plans, leading to improved decision-making and patient care.

**Security and Privacy Measures:** Ensuring the security and privacy of patient information is paramount in the healthcare domain, and the integration with EHR must adhere to stringent standards. The Health Insurance Portability and Accountability Act (HIPAA) sets forth guidelines for safeguarding sensitive patient data. Therefore, the integration must prioritize HIPAA compliance to guarantee the confidentiality and integrity of electronic health records.

* **HIPAA Compliance:** Adhering to HIPAA compliance involves implementing robust security measures, data encryption, access controls, and audit trails to protect electronic health records from unauthorized access or breaches. Compliance with HIPAA regulations not only safeguards patient information but also instills trust among healthcare professionals and patients in the secure handling of sensitive medical data.

# Improvement Result

##### **BERT (Bidirectional Encoder Representations from Transformers):**

BERT has gained significant popularity and has proven to be highly effective in various NLP tasks. In the context of text summarization, BERT's bidirectional architecture, which captures contextual information from both directions, can be advantageous. BERT's pre-training objective of masked language modeling enables it to capture deep contextualized representations, making it capable of generating informative and accurate summaries. BERT's ability to understand the nuances of language and capture important details makes it a strong contender for text summarization tasks

from transformers import BertTokenizer, BertModel

from summarizer import Summarizer

bert\_model = Summarizer()

bert\_summary = ''.join(bert\_model(body, min\_length=60))

print(bert\_summary)

#### **GPT (Generative Pre-trained Transformer):**

GPT is known for its impressive text generation capabilities and has been widely used in various language generation tasks. However, when it comes to text summarization, GPT faces certain limitations. Since GPT is autoregressive, it generates text sequentially, which can result in longer and more verbose summaries. GPT may struggle to produce concise and extractive summaries, which are essential for effective text summarization.

from transformers import pipeline

GPT3\_model = pipeline("text-generation", model="openai/gpt-3.5-turbo", tokenizer="openai/gpt-3.5-turbo")

response = GPT3\_model(body, max\_length=60)[0]['generated\_text']

print(response)

### **Comparing Models for Text Summarization: Factors to Consider**

When comparing these models for text summarization, certain factors come into play:

#### **Performance:**

BERT has been extensively evaluated and benchmarked on various NLP tasks, including text summarization, and has shown impressive performance. GPT's text generation capabilities may not be well-suited for concise summarization,performance in text summarization may require further exploration and evaluation.

#### **Training Data:**

The quality and diversity of training data play a crucial role in the performance of language models. BERT has been trained on large-scale corpora and is well-established in leveraging vast amounts of data. GPT can also benefit from large-scale training data, but the availability and specific datasets used may vary.

#### **Fine-tuning:**

Fine-tuning is an important step in optimizing language models for specific tasks. BERT's architecture and training objectives make it relatively straightforward to fine-tune for text summarization. GPT may require additional techniques and adaptations to achieve optimal performance for summarization.

#### **Resource Requirements:**

GPT is computationally intensive models, requiring substantial computational resources and time for training. BERT, on the other hand, is relatively less resource-intensive, making it more accessible and easier to deploy in practical applications.

#### **Latency :**

BERT have showcased impressive performance that resulted in lower latency. GPT's strength lies in text generation, but its autoregressive decoding mechanism could lead to increased response times, making it less suitable for latency critical applications.

#### **Reduced Dependency on Pre-processing:**

BERT often requires significant pre-processing steps like tokenization and data formatting to fit the fixed-length input requirements. LLMs, on the other hand, can handle more raw or unstructured text inputs, reducing the need for extensive pre-processing pipelines.

#### **Adaptability to Different Tasks:**

LLMs are pre-trained on vast amounts of data from diverse sources, allowing them to adapt to a wide range of natural language processing tasks. By providing task-specific prompts, LLMs can be fine-tuned to perform tasks such as text classification, question answering, summarization, or language translation, among others.

#### **Generation of Contextual Responses:**

LLMs are designed to generate coherent and contextually relevant responses based on the provided prompts. They have a better understanding of the context and can generate more natural and human-like responses compared to BERT, which primarily focuses on understanding the context of individual words or sentences.

##### References

1. G Qianqian Xie, Zheheng Luo, Benyou Wang, Sophia Ananiadou, “A Survey for Biomedical Text Summarization: From Pre-trained to Large Language Models”
2. Qianqian Xie,1 Edward J. Schenck,2,3 He S. Yang,4 Yong Chen,5 Yifan Peng,1 and Fei Wang, J,” Faithful AI in Medicine: A Systematic Review with Large Language Models and Beyond”
3. Prakhar Mishra, Ravi Theja Desetty,”NewAgeHealthWarriors at MEDIQA-Chat 2023 Task A: Summarizing Short Medical Conversation with Transformers”
4. Dave Van Veen\*, Cara Van Uden\*, Maayane Attias, Anuj Pareek, Christian Bluethgen, Malgorzata Polacin, Wah Chiu, Jean-Benoit Delbrouck, Juan Manuel Zambrano Chaves, Curtis P. Langlotz, Akshay S. Chaudhari, John Pauly,” RadAdapt: Radiology Report Summarization via Lightweight Domain Adaptation of Large Language Models”
5. “Evaluation of ChatGPT on Biomedical Tasks: A Zero-Shot Comparison with Fine-Tuned Generative Transformers”, Israt Jahan†, $, Md Tahmid Rahman Laskar‡, $, §, Chun Peng† , Jimmy Xiangji Huang‡, $