Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

A comprehensive evaluation of large Language models on benchmark biomedical text processing tasks

Israt Jahan [a,c], Md Tahmid Rahman Laskar [b,c,d], Chun Peng [a], Jimmy Xiangji Huang [b,c,*]

[a] Department of Biology, York University, Canada
[b] School of Information Technology, York University, Canada
[c] Information Retrieval and Knowledge Management Research Lab, York University, Canada
[d] Dialpad Inc., Canada

ARTICLE INFO

Keywords:
Large language models
ChatGPT
PaLM
LLaMA
Claude
Transformer
Natural language processing
LLM evaluation

ABSTRACT

Recently, **L**arge **L**anguage **M**odels (LLMs) have demonstrated impressive capability to solve a wide range of tasks. However, despite their success across various tasks, no prior work has investigated their capability in the biomedical domain yet. To this end, this paper aims to evaluate the performance of LLMs on benchmark biomedical tasks. For this purpose, a comprehensive evaluation of 4 popular LLMs in 6 diverse biomedical tasks across 26 datasets has been conducted. To the best of our knowledge, this is the first work that conducts an extensive evaluation and comparison of various LLMs in the biomedical domain. Interestingly, we find based on our evaluation that in biomedical datasets that have smaller training sets, zero-shot LLMs even outperform the current state-of-the-art models when they were fine-tuned only on the training set of these datasets. This suggests that pre-training on large text corpora makes LLMs quite specialized even in the biomedical domain. We also find that not a single LLM can outperform other LLMs in all tasks, with the performance of different LLMs may vary depending on the task. While their performance is still quite poor in comparison to the biomedical models that were fine-tuned on large training sets, our findings demonstrate that LLMs have the potential to be a valuable tool for various biomedical tasks that lack large annotated data.

1. Introduction

The rapid growth of language models [1] in the field of Natural Language Processing (NLP) in recent years has led to significant advancements in various domains, including the biomedical domain [2]. Although specialized models like BioBERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers for **Bio**medical Text Mining) [3], BioBART (**B**idirectional and **A**uto-**R**egressive **T**ransformers for the **Bio**medical Domain) [4], and BioGPT (**G**enerative **P**re-trained **T**ransformer for **Bio**medical Text Generation and Mining) [5] have shown promising results in the biomedical domain, they require fine-tuning[1] using domain-specific datasets. This fine-tuning process can be time-consuming due to the requirement of task-specific large annotated datasets. In contrast, zero-shot[2] learning [6] enables models to perform tasks without the need for fine-tuning on task-specific datasets.

Large **L**anguage **M**odels (LLMs) [7] are a class of natural language processing models that have been trained on vast amounts of textual data, making it possible to understand and generate human-like language. In recent years, LLMs such as ChatGPT[3] have demonstrated impressive performance on a range of language tasks, including text classification, question answering, and text summarization. One area where LLMs are not yet deeply investigated is the biomedical text processing and information retrieval domain. While there are vast amount of textual data available in the field of biomedicine, there still remains a scarcity of annotated datasets in this domain. Thus, it is difficult to build suitable models for biomedical tasks that lack large annotated datasets. In this regard, due to the strong zero-shot capabilities of LLMs across various tasks, LLM-powered automated tools can be useful for researchers and practitioners in the biomedical domain to find relevant information and extract insights from this vast corpus of unannotated data. However, despite being evaluated on various traditional NLP tasks, there is a lack of comprehensive studies that evaluate LLMs in the biomedical domain. To this end, this paper aims to evaluate LLMs across benchmark biomedical tasks.

* Correspondence to: 3048 Victor Phillip Dahdaleh Building (DB), York University, Canada.
E-mail addresses: israt18@yorku.ca (I. Jahan), tahmid20@yorku.ca (M.T.R. Laskar), cpeng@yorku.ca (C. Peng), jhuang@yorku.ca (J.X. Huang).

[1] Fine-tuning means providing good amount (e.g., thousands of samples) of training examples to re-train a pre-trained language model on a specific task.

[2] Zero-shot learning means asking a trained model to complete a task without providing any explicit examples of that particular task.

[3] https://chat.openai.com/

https://doi.org/10.1016/j.compbiomed.2024.108189
Received 11 October 2023; Received in revised form 14 February 2024; Accepted 18 February 2024
Available online 20 February 2024
0010-4825/© 2024 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

However, the evaluation of LLMs in the biomedical domain would require a proper understanding of the complex linguistic characteristics of biomedical texts. In addition, LLMs are sensitive to prompts [8,9]. Thus, for biomedical tasks, the effective construction of prompts is important to best utilize these LLMs in biomedical applications. Under these circumstances, domain-specific knowledge in the biomedical domain could play a pivotal role in improving the performance of LLMs in biomedical tasks. In this regard, we study how to effectively build prompts for LLMs to simulate common tasks in biomedical research, such as document classification, named entity recognition, relation extraction, text summarization, question answering, etc.

Since technologies in medicine and healthcare are critical, it is important to ensure rigorous evaluation before using LLMs in these domains. Thus, this paper will contribute to the understanding of the capabilities and limitations of LLMs in biomedical text processing and information retrieval. Moreover, with a comprehensive evaluation of various powerful LLMs, this paper would lead to the development of new tools and techniques for researchers in this field, which could pave the way to build new applications in healthcare and biomedicine via leveraging LLMs. The major contributions from this study are summarized below:

- A comprehensive evaluation of various LLMs in the biomedical domain, providing insights into their capabilities and limitations for various tasks. More specifically, this study investigates the zero-shot capabilities of LLMs in the Biomedical domain to address the lack of large annotated datasets in this domain.
- Construction of task-specific prompts by understanding the complex linguistic structure of biomedical texts. Our findings based on the extensive performance analysis of LLMs across various biomedical tasks will help researchers and practitioners when building LLM-based applications for the biomedical domain.
- To pave the way for future research on LLMs in the biomedical domain, we will release the code used for pre-processing and parsing of LLM-generated responses, alongside the data (the prompts constructed for LLMs and the LLM-generated responses) here: https://github.com/tahmedge/llm-eval-biomed.

2. Related work

There are a large number of studies on various biomedical tasks, such as biomedical image analysis [10–12], biomedical text processing [13,14], genomic sequence analysis [15,16], disease diagnosis [17], drug discovery [18–20], cancer research [21], vaccine development [22], etc. Biomedical text processing is closely related to these tasks as it serves as a critical component and enabler by providing automated methods for extracting information from the vast amount of textual data in the biomedical domain. In this section, we mainly review the existing state-of-the-art approaches for processing large amounts of biomedical textual data, that are the most related to our research. In the following, we first briefly review various language models used in recent years in the biomedical domain, followed by a brief review of the LLMs that have been studied in this paper.

2.1. Language models for the biomedical domain

In recent years, the effective utilization of transformer-based [23] NLP models like BERT [24] and GPT [25] have led to significant progress in the biomedical domain [3,26–30]. BERT leverages the encoder of the transformer architecture, while GPT leverages the decoder of the transformer. In addition to these models, sequence-to-sequence models like BART [31] that leverage both the encoder and the decoder of the transformer have also emerged as a powerful approach in various text generation tasks in the biomedical domain [4]. It has been observed that domain-specific pre-training of these models on the biomedical text corpora followed by fine-tuning on task-specific biomedical datasets have helped these models to achieve state-of-the-art performance in a variety of Biomedical NLP (BioNLP) tasks [32]. This led to the development of various language models for the biomedical domain, such as BioBERT [3], ClinicalBERT [26], BioBART [4], BioElectra [30], BioGPT [5], etc. However, one major limitation of using such fine-tuned models is that they require task-specific large annotated datasets, which is significantly less available in the BioNLP domain in comparison to the general NLP domain. In this regard, having a strong zero-shot model could potentially alleviate the need for large annotated datasets, as it could enable the model to perform well on tasks that it was not exclusively trained on.

2.2. Large language models

In recent years, large autoregressive decoder-based language models like GPT-3 [33] have demonstrated impressive few-shot learning capability. With the success of GPT-3 in few-shot scenarios, a new variant of GPT-3 called the InstructGPT model [34] has been proposed that leverages the reinforcement learning [35] from human feedback (RLHF) mechanism. The resulting InstructGPT models (in other words, GPT-3.5) are much better at following instructions than the original GPT-3 model, resulting in an impressive zero-shot performance across various tasks. ChatGPT[4] is the latest addition in the GPT-3.5 series models that additionally uses dialog-based instructional data during its training phase. Recently, more decoder-based LLMs such as PaLM[5] [36–38], Claude,[6] LLaMA[7] [39,40] etc. have been proposed that also achieve impressive performance in a wide range of tasks. All these LLMs including ChatGPT are first pre-trained on a large amount of textual data to predict the next token and then fine-tuned using a process called reinforcement learning from human feedback (RLHF) that leveraged both supervised learning and reinforcement learning techniques. The goal of RLHF was to improve the model's performance and ensure that it provided high-quality responses to user queries. The supervised learning phase of the RLHF process involved training the model on conversations in which human trainers played both sides: the user and the AI assistant. These conversations were collected from a variety of sources, including chat logs from customer service interactions, social media messages, and chatbots. The supervised learning phase aimed to train the model to produce high-quality responses that were contextually relevant to the user's query. Meanwhile, the reinforcement learning phase of the RLHF process aimed to further improve the model's performance by using human trainers to provide feedback on its responses. In this phase, human trainers ranked the responses that the model had created in a previous conversation. These rankings were used to create "reward models" that were used to fine-tune the model further by using several iterations of Proximal Policy Optimization (PPO) [35].

While these models have demonstrated strong performance in various NLP tasks [41–43], they have not been investigated in the biomedical domain yet. To this end, this paper aims to evaluate these powerful LLMs in the biomedical domain.

3. Biomedical tasks description

The biomedical text processing task refers to the use of computational techniques to analyze and extract information from textual data in the field of biomedicine. It can be defined as follows:

$$T : X \rightarrow Y \tag{1}$$

Here, X represents the input text for the given task T, and Y represents the output generated. In the following, the description of the benchmark biomedical text processing tasks that have been studied in this paper along with some examples are demonstrated.

[4] https://openai.com/blog/chatgpt

[5] https://ai.google/discover/palm2/

[6] https://www.claudeai.ai/

[7] https://ai.meta.com/blog/large-language-model-llama-meta-ai/

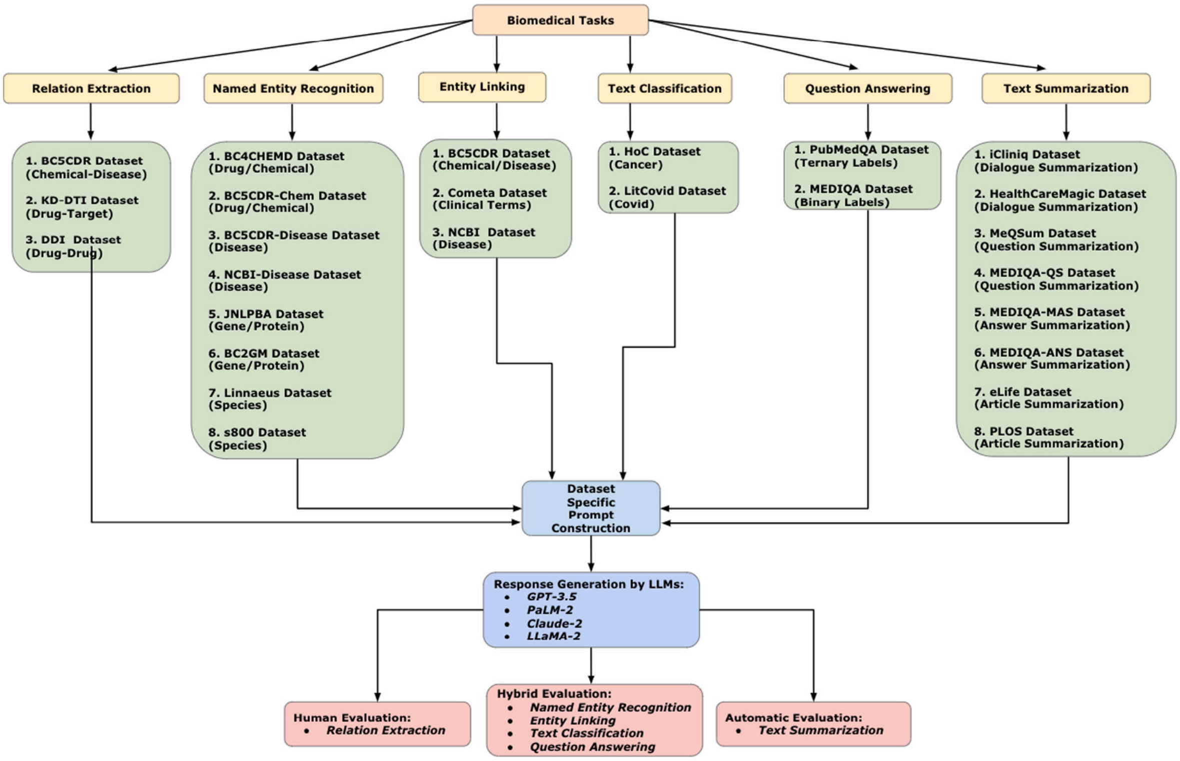

Biomedical Tasks
Relation Extraction
Named Entity Recognition
Entity Linking
Text Classification
Question Answering
Text Summarization
1. BC5CDR Dataset (Chemical-Disease)
2. KD-DTI Dataset (Drug-Target)
3. DDI Dataset (Drug-Drug)
1. BC4CHEMD Dataset (Drug/Chemical)
2. BC5CDR-Chem Dataset (Drug/Chemical)
3. BC5CDR-Disease Dataset (Disease)
4. NCBI-Disease Dataset (Disease)
5. JNLPBA Dataset (Gene/Protein)
6. BC2GM Dataset (Gene/Protein)
7. Linnaeus Dataset (Species)
8. s800 Dataset (Species)
1. BC5CDR Dataset (Chemical/Disease)
2. Cometa Dataset (Clinical Terms)
3. NCBI Dataset (Disease)
1. HoC Dataset (Cancer)
2. LitCovid Dataset (Covid)
1. PubMedQA Dataset (Ternary Labels)
2. MEDIQA Dataset (Binary Labels)
1. iCliniq Dataset (Dialogue Summarization)
2. HealthCareMagic Dataset (Dialogue Summarization)
3. MeQSum Dataset (Question Summarization)
4. MEDIQA-QS Dataset (Question Summarization)
5. MEDIQA-MAS Dataset (Answer Summarization)
6. MEDIQA-ANS Dataset (Answer Summarization)
7. eLife Dataset (Article Summarization)
8. PLOS Dataset (Article Summarization)
Dataset Specific Prompt Construction
Response Generation by LLMs:
GPT-3.5
PaLM-2
Claude-2
LLaMA-2
Human Evaluation:
Relation Extraction
Hybrid Evaluation:
Named Entity Recognition
Entity Linking
Text Classification
Question Answering
Automatic Evaluation:
Text Summarization

Fig. 1. An overview of our methodology to evaluate 6 biomedical tasks across 26 datasets in this paper. At first, we construct the prompt for each dataset. Then, we generate the response for each dataset using respective LLMs. Finally, depending on the task, we apply various evaluation techniques.

(i) Biomedical named entity recognition. Named Entity Recognition (NER) is the task of identifying named entities like person, location, organization, drug, disease, etc. in a given text [44]. In the case of biomedical NER, this task aims to extract the biomedical named entities, such as genes, proteins, diseases, chemicals, etc., from the literature to improve biomedical research.

Example: *The patient has been diagnosed with a rare form of cancer and is undergoing chemotherapy treatment with the drug Taxol.*

Expected NER classifications:

- NER (Disease): “rare form of cancer”.
- NER (Treatment): “chemotherapy”.
- NER (Drug): “Taxol”.

(ii) Biomedical relation extraction. The relation extraction task aims to extract relations between named entities in a given text [45]. In the biomedical relation extraction task, the aim is to analyze textual data by identifying which gene/variants are responsible for which diseases, which treatment/drug is effective for which disease, as well as identifying drug–drug interactions, etc.

Example: *The patient has been diagnosed with a rare form of cancer and is undergoing chemotherapy treatment with the drug Taxol.*

Expected Relation Extractions:

- Relation (Treatment of a Disease): “chemotherapy” is a treatment for “rare form of cancer”.
- Relation (Drug used in Treatment): “Taxol” is a drug used in “chemotherapy”.

(iii) Biomedical entity linking. The entity linking task focuses on linking named entities in a text to their corresponding entries in a knowledge base [46,47]. In the case of the biomedical entity linking task, it involves recognizing and linking biomedical named entities in unstructured text to their correct definitions, e.g., to the corresponding entries in structured knowledge bases or ontologies.

Example: *The patient has been diagnosed with a rare form of cancer and is undergoing chemotherapy treatment with the drug Taxol.*

Expected Entity Linking: A biomedical entity linking system may link the drug Taxol to the following link: *https://chemocare.com/druginfo/taxol*.

(iv) Biomedical text classification. For a given text, the goal of this task is to classify the text into a specific category. One example to classify a given sentence in one of the 10 hallmarks of cancer taxonomy has been demonstrated below:

Example: “Heterogeneity in DNA damage within the cell population was observed as a function of radiation dose.”

Expected Result: Genomic Instability and Mutation.

(v) Biomedical question answering. The biomedical question-answering task involves retrieving the relevant answer for the given question related to the biomedical literature, such as scientific articles, medical records, and clinical trials. This task is of great importance as it can help healthcare professionals, researchers, and patients access relevant information quickly and efficiently, which can have a significant impact on patient care, drug development, and medical research.

Example: *What is recommended for thalassemia patients ?*

- *Candidate Answer 1: Chemotherapy may be used to: Cure the cancer, shrink the cancer, and prevent the cancer from spreading.*
- *Candidate Answer 2: Regular blood transfusions can help provide the body with normal red blood cells containing normal hemoglobin.*

Expected Answer: The candidate answer 2 should be retrieved as a relevant answer [48,49].

(vi) Biomedical text summarization. The main purpose of the text summarization task is to generate a short concise summary of the given document [50]. The generation of short summaries of biomedical texts would help reduce the time spent reviewing lengthy electronic health records/patient queries in healthcare forums/doctor–patient conversations, resulting in improving the efficiency of the healthcare system.

Example: *Patient is a 62-year-old female with a medical history of hyperlipidemia, osteoarthritis, and previous cerebrovascular accident. She*

Table 1
Sample prompts in different named entity recognition (NER) datasets.

Dataset	Type	Data Split (Train/Valid/Test)	Prompt
BC2GM	NER (GENE/PROTEIN)	12574/2519/5038	Below, we provide a biomedical text: [TEXT] You need to identify the [ENTITY] type named entities in the above text. To identify the named entities, please tag each token of the given text in the ’BIO’ format as either: ’B’ or ’I’ or ’O’ The BIO format stands for Beginning, Inside, Outside. It provides a way to label individual tokens in a given text to indicate whether they are part of a named entity. In the BIO format, each token in a text is labeled with a tag that represents its role in a named entity. For our case, there are three possible tags: B: it indicates that the token is the beginning of the [ENTITY] type named entity (i.e., the first token of a [ENTITY] type named entity). I: it indicates the token is inside a [ENTITY] type named entity (i.e., any token other than the first token of a [ENTITY] type named entity). O: it indicates that the token is outside any named entity. In other words, it is not part of any named entity. Below, each token of the biomedical text is provided (separated by new line). Now please assign the correct tag to each token. Return your result for each token in a newline in the following format -> token: assigned_tag: [LIST OF LINE SEPARATED TOKENS]
BC4CHEMD	NER (DRUG/CHEMICAL)	30682/30639/26364	
BC5CDR-CHEM	NER (DRUG/CHEMICAL)	4560/4581/4797	
BC5CDR-Disease	NER (DISEASE)	4560/4581/4797	
JNLPBA	NER (GENE/PROTEIN)	14690/3856/3856	
LINNAEUS	NER (SPECIES)	11935/4078/7142	
NCBI-Disease	NER (DISEASE)	5424/923/940	
s800	NER (SPECIES)	5733/830/1630	

Table 2
Sample prompts in different relation extraction datasets.

Dataset	Type	Data Split (Train/Valid/Test)	Prompt
BC5CDR	Chemical-Disease Relation Extraction	500/500/500	Identify each pair of drugs and the drug-induced side-effects (e.g., diseases) in the following passage: [PASSAGE]
KD-DTI	Drug-Target Relation Extraction	12K/1K/1.3K	Identify the drug-target interactions in the following passage (along with the interaction type among the following: ’inhibitor’, ’agonist’, ’modulator’, ’activator’, ’blocker’, ’inducer’, ’antagonist’, ’cleavage’, ’disruption’, ’intercalation’, ’inactivator’, ’bind’, ’binder’, ’partial agonist’, ’cofactor’, ’substrate’, ’ligand’, ’chelator’, ’downregulator’, ’other’, ’antibody’, ’other/unknown’): [PASSAGE]
DDI	Drug–Drug Relation Extraction	664/50/191	Identify the pairs of drug–drug interactions in the passage given below based on one of the following interaction types: (i) mechanism: this type is used to identify drug–drug interactions that are described by their pharmacokinetic mechanism. (ii) effect: this type is used to identify drug–drug interactions describing an effect. (iii) advice: this type is used when a recommendation or advice regarding a drug–drug interaction is given. (iv) int: this type is used when a drug–drug interaction appears in the text without providing any additional information. [PASSAGE]

Table 3
Sample prompts in different entity linking datasets.

Dataset	Type	Data split (Train/Valid/Test)	Prompt
BC5CDR	Entity Linking (DISEASE/CHEMICAL)	9285/9515/9654	[TEXT_S <START> ENTITY <END> TEXT_E] In the biomedical text given above, what does the entity between the START and the END token refer to?
COMETA	Entity Linking (CLINICAL TERMS)	13489/2176/4350	
NCBI	Entity Linking (DISEASE)	5784/787/960	

presented with sudden onset of dizziness and palpitations that began a day ago. An electrocardiogram was immediately conducted, which indicated the presence of atrial fibrillation. She was promptly hospitalized for monitoring and commenced on anticoagulation therapy with warfarin and rate-controlling medications like beta-blockers.

***Expected Summary:** A 62-year-old female with a history of hyperlipidemia, osteoarthritis, and a previous cerebrovascular accident experienced sudden dizziness and palpitations. An ECG confirmed atrial fibrillation, leading to her hospitalization and treatment with warfarin and beta-blockers.*

4. Methodology

In this section, we first present our methodology on how we design the prompts for different tasks, followed by describing the LLMs that have been studied in this paper. Afterward, the evaluation pipeline has been demonstrated. An overview of our methodology is also shown in Fig. 1.

4.1. Prompt design

For a given test sample X, we first prepare a task instruction T. Then, we concatenate the test sample X with the task instruction T to construct the prompt P. Afterward, the prompt P is given as input to generate the response R. Below, the prompt P that has been constructed for each task depending on the respective dataset has been demonstrated.

(i) NER. For NER, prompts are designed to identify the biomedical named entities in a given text in the BIO format. In our prompts, the description of the BIO format is also added along with the task instructions. For NER, we use the BC2GM [51] and JNLPBA [52] datasets for gene/protein entity recognition, BC4CHEMD [53] and BC5CDR-CHEM [54] for drug/chemical entity recognition, BC5CDR-Disease [54] and NCBI-Disease [55] for disease type entity recognition, LINNAEUS [56] and s800 [57] for species type entity recognition. The prompts for this task are shown in Table 1.

Table 4
Sample prompts in different text classification datasets.

Dataset	Type	Data split (Train/Valid/Test)	Prompt
HoC	Text Classification	9972/4947/4947	The 10 hallmarks of cancer taxonomy with their definitions are given below: (i) Sustaining proliferative signaling: Cancer cells can initiate and maintain continuous cell division by producing their own growth factors or by altering the sensitivity of receptors to growth factors. (ii) Evading growth suppressors: Cancer cells can bypass the normal cellular mechanisms that limit cell division and growth, such as the inactivation of tumor suppressor genes and/or insensitivity to antigrowth signals. (iii) Resisting cell death: Cancer cells develop resistance to apoptosis, the programmed cell death process, which allows them to survive and continue dividing. (iv) Enabling replicative immortality: Cancer cells can extend their ability to divide indefinitely by maintaining the length of telomeres, the protective end caps on chromosomes. (v) Inducing angiogenesis: Cancer cells stimulate the growth of new blood vessels, providing the necessary nutrients and oxygen to support their rapid growth. (vi) Activating invasion and metastasis: Cancer cells can invade surrounding tissues and migrate to distant sites in the body, forming secondary tumors called metastases. (vii) Deregulating cellular energetic metabolism: Cancer cells rewire their metabolism to support rapid cell division and growth, often relying more on glycolysis even in the presence of oxygen (a phenomenon known as the Warburg effect). (viii) Avoiding immune destruction: Cancer cells can avoid detection and elimination by the immune system through various mechanisms, such as downregulating cell surface markers or producing immunosuppressive signals. (ix) Tumor promoting inflammation: Chronic inflammation can promote the development and progression of cancer by supplying growth factors, survival signals, and other molecules that facilitate cancer cell proliferation and survival. (x) Genome instability and mutation: Cancer cells exhibit increased genomic instability, leading to a higher mutation rate, which in turn drives the initiation and progression of cancer. Classify the sentence given below in one of the above 10 hallmarks of cancer taxonomy (if relevant). If cannot be classified, answer as "empty": [SENTENCE]
LitCovid	Text Classification	16126/2305/4607	Choose the most appropriate topic(s) for the biomedical article on covid-19 given below from the following options: (i) Prevention, (ii) Treatment, (iii) Diagnosis, (iv) Mechanism, (v) Case Report, (vi) Transmission, (vii) Forecasting, and (viii) General. [ARTICLE]

(ii) Relation extraction. To identify the possible relation between entities mentioned in a given text, the prompts are designed depending on the dataset. For this purpose, we construct prompts for chemical-disease-relation in the BC5CDR dataset [54], drug-target-interaction in the KD-DTI dataset [58], and drug–drug-interaction in the DDI dataset [59]. The prompts used for these datasets are demonstrated in Table 2.

(iii) Entity linking. To identify whether LLMs can link named entities to their correct definitions based on their pre-training knowledge, we follow the work of Yuan et al. [60] for the generative entity linking task by asking LLMs to identify the correct concept names for the named entities. For evaluation, the BC5CDR [54] dataset for the entity linking of disease/chemical type named entities, the NCBI [55] dataset to link diseases, and the COMETA [61] dataset to link clinical terms have been used. The sample prompts for this task are shown in Table 3.

(iv) Text classification. The goal of this task is to classify the type of the given text. In this paper, we use two datasets: (i) the HoC (the Hallmarks of Cancer corpus) dataset [62], and (ii) the LitCovid dataset [63]. The HoC dataset consists of 1580 PubMed abstracts where the goal is to annotate each sentence in the given abstract in one of the 10 currently known hallmarks of cancer. Whereas in the LitCovid dataset, each article is required to be classified in one (or more) of the following 8 categories: Prevention, Treatment, Diagnosis, Mechanism, Case Report, Transmission, Forecasting, and General. Our prompts for these text classification datasets are shown in Table 4.

(v) Question answering. For the question-answering task, we also evaluate the performance of LLMs on multiple datasets: (i) the PubMedQA dataset [64], and (ii) the MEDIQA-2019 dataset [48]. In the PubmedQA dataset, the question, the reference context, and the answer are given as input to the LLMs to determine whether the answer to the given question can be inferred from the provided reference context with LLMs being prompted to reply either as *yes*, *no*, or *maybe*, as required by the task. In the MEDIQA-2019 dataset, the LLMs are asked to determine whether the retrieved answer for the given question is relevant or not [65]. The prompts for this task are shown in Table 5.

(vi) Text summarization. The biomedical text summarization task requires the generation of a concise summary of the given biomedical text. To this end, the LLMs are evaluated across a wide range of diverse biomedical summarization tasks, such as healthcare question summarization (*MeQSum* [66] and *MEDIQA-QS* [67] datasets), medical answer summarization (*MEDIQA-ANS* [68] and *MEDIQA-MAS* [67] datasets), and doctor–patient dialogue summarization (*iCliniq* and *HealthCareMagic* datasets [69,70]) to generate short queries for healthcare forums describing patient's medical conditions. In addition, we use various datasets for biomedical literature summarization [71,72], such as the Biomedical Text Lay Summarization shared task 2023 (BioLaySumm-2023) datasets [73]. For BioLaySumm-2023, since the gold reference summaries of the test sets are not publicly available as of the writing of this paper, the respective validation sets are used for evaluation. The sample prompts in the summarization datasets are shown in Table 6.

4.2. Models

In the following, we describe the 4 popular LLMs that we evaluate in benchmark biomedical datasets and tasks in this paper.

(i) GPT-3.5. GPT-3.5 is an auto-regressive language model based on the transformer [23] architecture that was pre-trained on a vast amount of textual data via supervised learning alongside reinforcement learning with human feedback. The backbone model behind the first version of ChatGPT was also GPT-3.5, and it is currently one of the base models, behind OpenAI's ChatGPT, alongside GPT-4. The initial training data for GPT-3.5 was obtained from a large corpus of text data that was crawled from the internet. This corpus included a wide range of

Table 5
Sample prompts in different question answering datasets.

Dataset	Type	Data split (Train/Valid/Test)	Prompt
PubMedQA	Question Answering	450/50/500	For the question, the reference context, and the answer given below, is it possible to infer the answer for that question from the reference context? Only reply as either Yes or No or Maybe. Question: [QUESTION] Reference context: [REFERENCE CONTEXT] Answer: [ANSWER]
MEDIQA-2019	Question Answering	1701/234/1107	A retrieved answer for the following question is given below. Identify whether the retrieved answer is relevant to the question or not. Answer as 1 if relevant, otherwise answer as 0. Question: [QUESTION] Retrieved Answer: [TEXT]

Table 6
Sample prompts in different text summarization tasks.

Dataset	Type	Data split (Train/Valid/Test)	Prompt
iCliniq	Dialogue Summarization	24851/3105/3108	Write a very short and concise one line summary of the following dialogue as an informal question in a healthcare forum: [DIALOGUE]
HealthCare Magic	Dialogue Summarization	181122/22641/22642	Write a very short and concise one line summary of the following dialogue as a question in a healthcare forum: [DIALOGUE]
MeQSum	Question Summarization	500/-/500	Rewrite the following question in a short and concise form: [QUESTION]
MEDIQA-QS	Question Summarization	-/50/100	Rewrite the following question in a short and concise form: [QUESTION]
MEDIQA-MAS	Answer Summarization	-/50/80	For the following question, some relevant answers are given below. Please write down a short concise answer by summarizing the given answers. Question: [QUESTION] Answer 1: [ANSWER1] Answer 2: [ANSWER2]
MEDIQA-ANS	Answer Summarization	-/-/552	Write a very short and concise summary of the following article based on the question given below: [QUESTION] [ARTICLE]
BioLaySumm-2023 (PLOS)	Lay Summarization	24773/1376/142	Write down a readable summary of the following biomedical article using less technical terminology (e.g., lay summary) such that it can be understandable for non-expert audiences: [ABSTRACT + ARTICLE]
BioLaySumm-2023 (eLife)	Lay Summarization	4346/241/142	Write down a readable summary of the following biomedical article using less technical terminology (e.g., lay summary) such that it can be understandable for non-expert audiences: [ABSTRACT + ARTICLE]
BioLaySumm-2023 (PLOS)	Readability-controlled Summarization (Lay Summary)	24773/1376/142	Write down a readable summary of the following biomedical article using less technical terminology (e.g., lay summary) such that it can be understandable for non-expert audiences: [ARTICLE]
BioLaySumm-2023 (PLOS)	Readability-controlled Summarization (Abstract)	24773/1376/142	Write down the abstract of the following biomedical article: [ARTICLE]

publicly available text, including articles, books, and websites. Additionally, OpenAI collected data from GPT-3 users to train and fine-tune the model further [41,74]. In this work, we used the OpenAI API for the *gpt-3.5-turbo-0613*[8] model for GPT-3.5.

(ii) PaLM-2. PaLM-2 [37] is also a transformer-based language model that exhibits enhanced multilingual and reasoning capabilities, along with improved computing efficiency. It is the base model behind Google's BARD,[9] which is a competitor to OpenAI's ChatGPT. The computational efficiency in PaLM-2 is achieved by scaling the model size and the training dataset size in proportion to each other. This new technique makes PaLM-2 smaller than its predecessor, PaLM-1, while achieving better performance, including faster inference, fewer parameters to serve, and a lower serving cost. It is trained using a mixture of objectives, allowing it to learn various aspects of language and reasoning across a diverse set of tasks and capabilities, making it a powerful tool for various applications. In this work, we used the *text-bison@001* model in Google's Vertex AI[10] API for PaLM-2.

(iii) Claude-2. Claude-2 is also a general-purpose LLM based on the transformer architecture. It was developed by Anthropic[11] and is a successor of Claude-1. Similar to other large models, it is trained via unsupervised pre-training, supervised fine-tuning, and reinforcement learning with human feedback. Internal red-teaming evaluation by Anthropic shows that Claude is more harmless and less likely to produce

[8] https://platform.openai.com/docs/models/gpt-3-5

[9] https://bard.google.com/

[10] https://cloud.google.com/vertex-ai/docs/generative-ai/model-reference/text

[11] https://www.anthropic.com/index/claude-2

offensive or dangerous output. Experimental evaluation of Claude-1 and Claude-2 demonstrates that Claude-2 achieves much better performance than Claude-1 across various tasks. Thus, we also utilize Claude-2 in this work via leveraging Anthropic's *claude-2* API.

(iv) LLaMA-2. LLaMA-2 [40] is a recently proposed LLM by Meta.[12] One major advantage of LLaMA-2 over the previously mentioned LLMs is that it is also open-sourced. While another open-sourced version of LLaMA: the LLaMA-1 [39] model was released prior to the release of LLaMA-2, the LLaMA-1 model was only allowed for non-commercial usage. On the contrary, the recently proposed LLaMA-2 not only allows commercial usage, but also outperforms its earlier open-sourced version LLaMA-1 across a wide range of tasks. This makes LLaMA-2 a breakthrough model in both academia and industry. Similar to other LLMs, LLaMA-2 is also trained via unsupervised pre-training, supervised fine-tuning, and reinforcement learning with human feedback. Note that the LLaMA-2 model has been released in various sizes: 7B, 13B, and 70B. While the 70B model has achieved the best performance across various benchmarks, it requires very high computational resources. On the other hand, although the 7B model requires less computational resources, it achieves poorer performance in comparison to the 13B and 70B models. Considering the performance and cost trade-off, we used the LLaMA-2-13B[13] model in this work.

4.3. Evaluation pipeline

Since LLMs usually generate human-like responses that may sometimes contain unnecessary information while not in a specific format, some tasks are very challenging to evaluate without any human intervention. For instance, in tasks like Relation Extraction, there can be multiple answers. Thus, it would be very difficult to automatically evaluate the performance of LLMs by comparing their response with the gold labels using just an evaluation script. Thus, in this paper, to ensure high-quality evaluation, we follow the work of Laskar et al. [75], where they design different settings for the evaluation of LLMs for different tasks:

i. **Automatic Evaluation:** Where they evaluate some tasks, such as text summarization via leveraging automatic evaluation scripts.
ii. **Human Evaluation:** Where they evaluate some discriminative tasks solely by humans, which cannot be evaluated directly based on automatic evaluation scripts.
iii. **Hybrid (Human + Automatic) Evaluation:** Where they evaluate some tasks via leveraging both human intervention alongside evaluation scripts. More specifically, this is done by first applying evaluation scripts on the dataset to parse the results from the LLM-generated response, followed by utilizing human intervention if solely depending on the evaluation script cannot parse the results in the expected format.

For discriminative tasks, where parsing of the results from the generated response is required for evaluation, we follow the work of Laskar et al. [75] and design an evaluation script for the respective dataset to first parse the results and then compare the parsed results with the gold labels. Subsequently, any samples where the script could not parse the result properly were manually reviewed by the human annotators. For NER, Entity Linking, Text Classification, and Question Answering, we evaluate the performance by leveraging this technique (denoted as *hybrid evaluation*). However, for relation extraction, human intervention is necessary since parsing scripts cannot properly identify the relations found in the generative responses. Thus, for relation extraction, all LLM-generated responses were manually evaluated by humans. This technique of solely utilizing humans to evaluate LLM-generated response when parsing is not possible was also used in recent literature [9,75]. In our human evaluation, at least two annotators compared the LLM-generated response against the gold labels. Any disagreements were resolved based on discussions between the annotators.

For generative tasks, such as summarization, where the full response generated by LLMs can be used for evaluation instead of parsing the response, we evaluate using automatic evaluation metrics (e.g., ROUGE or BERTScore).

5. Experiments

5.1. Evaluation metrics

We use different evaluation metrics for different tasks to ensure a fair comparison of different LLMs with prior state-of-the-art results. For this purpose, the standard evaluation metrics that are used in the literature for benchmarking the performance of different models are selected. Thus, for the relation extraction and named entity recognition tasks, Precision, Recall, and F1 metrics are used, while for entity linking, the Recall@1 metric is used. For Summarization, the ROUGE [76] and the BERTScore [77] metrics are used. For question answering and text classification, metrics like Accuracy and F1 are used.

5.2. Baselines

To compare the performance of the zero-shot LLMs, the current state-of-the-art fine-tuned models are used as the baselines. These baseline models are described below.

(i) BioGPT. The backbone of BioGPT [5] is GPT-2 [25], which is a decoder of the transformer [23]. The BioGPT model was trained over PubMed titles and abstracts via leveraging the standard language modeling task. We use the fine-tuned BioGPT models as the baseline for all datasets in the relation extraction task, HoC dataset in the text classification task, and the PubMedQA[14] dataset for the question-answering tasks.

(ii) BioBART. It is a sequence-to-sequence model based on the BART [31] architecture where the pre-training process involves reconstructing corrupted input sequences. The main difference between BioBART [4] and BART is that the former was pre-trained over PubMed abstracts to make it suitable for the biomedical domain tasks. The fine-tuned BioBART model was used as the baseline in all the entity linking datasets and the following biomedical summarization tasks: Dialogue Summarization, Question Summarization, and Answer Summarization.

(iii) BioBERT. It is a domain-specific language representation model [3] based on the BERT [24] architecture that was additionally pre-trained on large-scale biomedical corpora (PubMed and PMC abstracts). The fine-tuned BioBERT model achieved state-of-the-art performance across different biomedical NER datasets and so it was used as the baseline for all NER datasets in this paper. In addition, it was used as the baseline in the LitCovid dataset for text classification.

(iv) ALBERT with disease knowledge infused. The ALBERT [78] model is a variant of the BERT [24] language model which requires lower memory consumption and a new self-supervised loss function. He at al., [49] extends its training mechanism by additionally training ALBERT on 14 K biomedical texts in a question-answering fashion via infusing disease knowledge which led to the state-of-the-art performance in the MediQA-2019 dataset. The LLMs are compared with this *disease knowledge infused* version of the ALBERT model in this work.

[12] https://ai.meta.com/llama/

[13] We used the following version of LLaMA-2-13B: https://huggingface.co/meta-llama/Llama-2-13b-chat-hf, which achieves improved factual correctness than its based version. As we are benchmarking LLMs in the biomedical domain, selecting a more faithful model is prioritized.

[14] In PubMedQA, BioGPT was additionally fine-tuned on more than 270 K instances.

Table 7
Performance on relation extraction datasets. All SOTA results are taken from the BioGPT [5] model.

Model	Dataset								
	BC5CDR			KD-DTI			DDI		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
GPT-3.5	30.62	73.85	43.29	19.19	66.02	29.74	**47.11**	**45.77**	**46.43**
PaLM-2	**51.61**	57.30	**54.30**	**40.21**	36.82	**38.44**	35.47	16.48	22.50
Claude-2	44.04	67.73	53.37	17.99	**72.73**	28.84	39.27	46.60	42.62
LLaMA-2-13b	39.54	**81.66**	53.28	15.14	60.48	24.21	22.58	25.67	24.03
State-of-the-Art (SOTA)	49.52	43.25	46.17	40.00	39.72	38.42	41.70	44.75	40.76

Table 8
Performance on Text Classification, Question Answering (QA), and Entity Linking datasets. The SOTA results for HoC and PubMedQA are taken from the BioGPT [5] model, while we take the SOTA results from Gutiérrez et al. [83] and He et al. [49] for LitCovid and MediQA-2019, respectively. Note that all SOTA results for Entity Linking are taken from the BioBART [4] model.

Model	Text classification dataset		Question answering dataset		Entity linking dataset		
	HoC	LitCovid	PubMedQA	MediQA-2019	BC5CDR	Cometa	NCBI
	F1	F1	Accuracy	Accuracy	Recall @ 1	Recall @ 1	Recall @ 1
GPT-3.5	59.26	29.63	54.40	**73.26**	54.90	43.45	52.19
PaLM-2	**61.03**	**37.50**	59.60	52.12	52.14	48.76	38.44
Claude-2	34.93	7.60	57.20	65.13	**78.01**	**53.29**	**70.21**
LLaMA-2-13b	41.82	11.34	**61.40**	56.01	66.52	40.67	59.17
State-of-the-Art (SOTA)	**85.12**	**86.20**	**78.20**	**79.49**	**93.26**	**81.77**	**89.90**

(v) FLAN-T5-XL:. FLAN-T5 [79] is an extension of the T5 [80] model. The T5 model treats each tasks as a sequence to sequence problem. While the architecture of FLAN-T5 is similar to the original T5 model, it leverage instruction fine-tuning instead of traditional fine-tuning. The FLAN-T5-XL that achieves state-of-the-art performance in the Biomedical Lay Summarization task is used as the baseline in the eLife and the PLOS datasets to compare LLMs in biomedical lay summarization.

(vi) PRIMERA. It is a pre-trained model [81] designed to enhance multi-document summarization. It proposes a new pre-training strategy for multi-document summarization by leveraging the longformer-encoder–decoder [82] for pre-training. In this work, the fine-tuned PRIMERA model is used as the baseline in the Readability-Controlled Summarization task since it is the current state-of-the-art in this task.

5.3. Results

In this section, the results for LLMs in various tasks are presented. At first, we present our results in the Relation Extraction task where we utilize *human evaluation*. Then, we demonstrate our findings in Text Classification, Question Answering, Entity Linking, and NER datasets where *hybrid evaluation* is conducted. Finally, we present our findings in the Summarization datasets where *automatic evaluation* is utilized.

(i) Relation extraction. We compare the performance of LLMs with the current state-of-the-art fine-tuned BioGPT [5] model across 3 datasets for the relation extraction task. The LLM generated responses in the relation extraction task are computed based on ***Human Evaluation***. From the results presented in Table 7, we find that in the BC5CDR dataset, while LLaMA-2 achieves the highest recall, PaLM-2 performs the best in terms of Precision and F1. Meanwhile, in terms of F1, the zero-shot PaLM-2, Claude-2, and LLaMA-2 model even outperform the prior state-of-the-art fine-tuned BioGPT in this dataset, with an improvement of 17.61% by the best performing PaLM-2. In the KD-DTI dataset, though GPT-3.5 and Claude-2 achieve high recall, their overall F1-score was quite lower than BioGPT and PaLM-2. Meanwhile, zero-shot PaLM-2 again performs much better while achieving almost similar performance in comparison to the fine-tuned BioGPT in terms of the F1 score. In the DDI dataset, GPT-3.5 achieves state-of-the-performance across all three metrics (Precision, Recall, and F1), followed by Claude-2. Since in the DDI dataset, there are only 4 types of labels, more descriptive prompts are used in this dataset (e.g., providing the definition of different interaction types), which helped GPT-3.5 and Claude-2 to achieve better performance. However, more descriptive prompts were not helpful for PaLM-2 in this dataset. Nonetheless, the impressive results achieved by LLMs in comparison to the prior state-of-the-art results in BC5CDR and DDI datasets demonstrate that in datasets having smaller training sets (both datasets have less than 1000 training samples), LLMs are more effective than even fine-tuned models. Meanwhile, in the KD-DTI dataset that has about 12 K training samples, most zero-shot LLMs still achieve comparable performance, with PaLM-2 slightly outperforming the state-of-the-art result. More interestingly, while other LLMs achieve quite poor precision scores in the KD-DTI dataset, PaLM-2 even outperforms the current state-of-the-art result in terms of precision. However, based on paired t-test with $p \leq .05$, the performance difference between the LLMs and the current fine-tuned SOTA models in terms of F1 is **not statistically significant**.

(ii) Text classification. In terms of Text Classification (see Table 8), the LLM generated responses are evaluated based on ***Hybrid Evaluation***. In comparison to the current state-of-the-art models fine-tuned on the respective datasets (BioGPT [5] in HoC and BioBERT [3] in LitCovid), it is evident that the zero-shot LLMs perform very poorly in comparison to the state-of-the-art fine-tuned baselines in both datasets. In particular, the performance of Claude-2 was much poorer than other LLMs. Among LLMs, GPT-3.5 and PaLM-2 are generally better, with PaLM-2 being the best performing LLM in both the HoC dataset and the LitCovid dataset. The difference in performance between the best performing PaLM-2 and the worst performing Claude-2 is also **statistically significant**, based on paired t-test, with $p \leq .05$.

We also investigate the effect of prompt tuning by evaluating two new prompts that are less descriptive, i.e., without giving definitions of the HoC classes, or without naming the HoC classes. Below our findings for GPT-3.5 based on prompt variations are demonstrated:

(i) Prompting with only the name of each HoC class is given without any definitions, drops the F1 score to 46.93.

(ii) Prompting without explicitly mentioning the name of 10 HoC classes, drops F1 to 38.20.

This indicates that for classification tasks, descriptive prompts are very helpful in improving the performance of LLMs (see Section 5.4.1 for more details).

(iii) Question answering. For question answering, we evaluate the performance based on ***Hybrid Evaluation*** on two datasets (see Table 8).

In terms of the question-answering task in the PubMedQA dataset, we find that the performance of all LLMs is much lower than the current

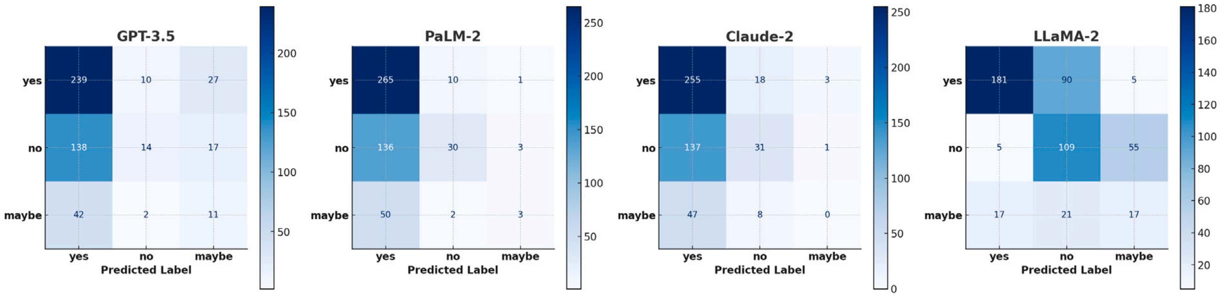


Fig. 2. Confusion Matrix for different models in the PubMedQA dataset.

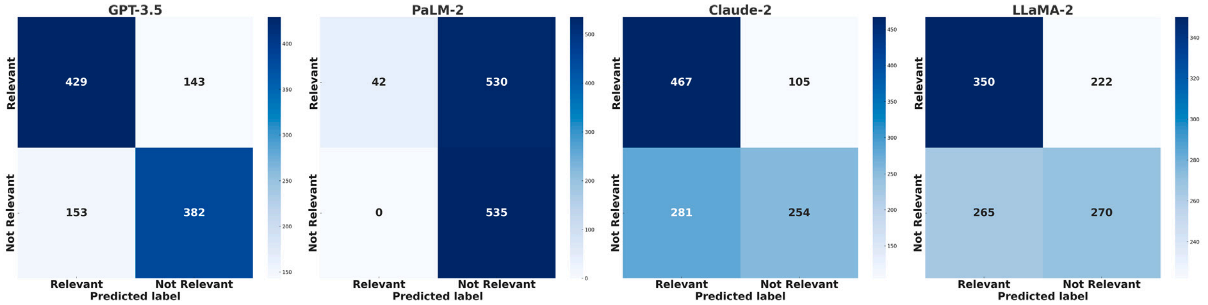


Fig. 3. Confusion Matrix for different models in the MediQA-2019 dataset.

state-of-the-art BioGPT model. It should be noted that the BioGPT [5] model which achieves the state-of-the-art result in PubmedQA was additionally trained on the PQA-A (211K instances) and PQA-U (61K instances) splits of the PubmedQA dataset (along with the PQA-L split which is the dedicated training set of this dataset). While comparing the performance of the closed-source LLMs (GPT-3.5, PaLM-2, Claude-2), we find that they perform almost similarly, with none of them achieving more than 60% accuracy. More interestingly, none of these closed-source LLMs could outperform the LLaMA-2 model that achieves the best performance among LLMs in this dataset. This is an interesting finding since the LLaMA-2 only has 13B parameters, which is much smaller than the closed-source LLMs. To further investigate how LLaMA-2 achieves superior performance in this dataset, we present the confusion matrix using a heatmap based on the prediction made by different LLMs in Fig. 2. From the heatmap, we find that all LLMs except LLaMA-2 make mistakes while predicting the "no" type label, as in most cases the LLMs (GPT-3.5, PaLM-2, Claude-2) ended up predicted with the "yes" type label instead, leading to an overall poor accuracy.

In terms of the question-answering task in the MediQA-2019 dataset, we find that the accuracy from the PubMedQA dataset is increased for GPT-3.5 and Claude-2, while being decreased for the LLaMA-2 and PaLM-2; with the zero-shot GPT-3.5 achieving the best accuracy (73.26). The performance of GPT-3.5 is comparable to the current state-of-the-art accuracy of 79.49 [49] by the ALBERT model [78] which was additionally trained in question-answering style on 14 K biomedical texts consisting of disease-related knowledge followed by being fine-tuned on the MediQA-2019 dataset. To further investigate the performance of LLMs in this dataset, we show the confusion matrix in Fig. 3 to find that the best performing LLM in the MediQA-2019 dataset, the GPT-3.5 model was able to classify the Relevant and Not Relevant labels more accurately than other LLMs. Moreover, the reason behind PaLM-2 being the worst performer in this dataset is due to the fact that it predicts most instances as Not Relevant. Paired t-test with $p \leq .05$ demonstrates that the performance difference between the LLMs in question answering is **not statistically significant.**

(iv) Entity linking. All the entity linking datasets are evaluated based on the ***Hybrid Evaluation*** technique. For entity linking, we find from Table 8 that Claude-2 outperforms all other LLMs in all three entity linking datasets: BC5CDR, Cometa, and NCBI. In BC5CDR and NCBI, while LLaMA-2 is the second best performing model; the PaLM-2 is found to be the second best performer in the Cometa dataset. Nonetheless, the performance of the second best performing models is still quite below in comparison to the Claude-2 model. This finding suggests that Claude-2 is more useful than other models in biomedical entity linking tasks by effectively retrieving the correct definition from its pre-training knowledge, although its performance is still much below compared to the current fine-tuned SOTA models, which is also **statistically significant**, based on paired t-test with $p \leq .05$.

(v) NER. Similar to Entity Linking, we also conduct ***Hybrid Evaluation*** for NER and find from Table 9 that Claude-2 again outperforms the rest other LLMs across all NER datasets (also in terms of all evaluation metrics: *Precision*, *Recall*, and *F1*). However, the performance of all LLMs is significantly lower than the current SOTA results (based on paired t-test, this difference in performance is **statistically significant**, with $p \leq .05$), with the performance of LLaMA-2 being the poorest. Such limitations of zero-shot LLMs in NER have also been observed in datasets from the general NLP domain [75]. These findings give a strong indication that generative LLMs need further improvement on sequence labeling tasks like NER using the traditional BIO formatting.

(vi) Summarization. We present the results on the following summarization datasets: *Dialogue Summarization*, *Question Summarization*, and *Answer Summarization* in Table 10 and compare with BioBART [4]. For evaluation [86], we use the following two ***Automatic Evaluation*** metrics: (i) the widely used ROUGE [76] metric, and (ii) the BERTScore [77] metric. For BERTScore, we use the RoBERTa-Large [87] model for implementation. For all LLMs, the input context length of 2000 words has been used.

We observe that in terms of the ROUGE metric, all LLMs perform much worse than BioBART in datasets that have dedicated training sets, such as iCliniq, HealthCareMagic, and MeQSum. Meanwhile, they perform on par with BioBART in the MEDIQA-QS dataset. Among LLMs, in general, GPT-3.5 is found to be the best performer in these datasets. More importantly, GPT-3.5 outperforms BioBART in both MEDIQA-ANS and MEDIQA-MAS datasets. Note that MEDIQA-ANS, MEDIQA-MAS, and MEDIQA-QS datasets do not have any dedicated

Table 9
Performance on Named Entity Recognition datasets. SOTA results are from the BioBERT [3] model. Here, 'Precision' and 'Recall' are denoted by 'P' and 'R', respectively.

Dataset	Model														
	GPT-3.5			PaLM-2			Claude-2			LLaMA-2-13b			SOTA		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BC2GM	23.07	52.19	31.99	24.65	48.77	32.75	**31.95**	**55.10**	**40.45**	3.39	24.11	5.95	**84.32**	**85.12**	**84.72**
BC4CHEMD	17.33	52.08	26.01	18.27	44.09	25.83	**26.37**	**52.83**	**35.18**	3.67	35.05	6.64	**92.80**	**91.92**	**92.36**
BC5CDR-chem	29.93	66.30	41.25	37.93	65.63	48.08	**49.99**	**69.23**	**58.05**	6.98	48.41	12.21	**93.68**	**93.26**	**93.47**
BC5CDR-disease	23.37	52.08	32.26	26.56	46.16	33.72	**47.06**	**53.62**	**50.13**	3.16	27.98	5.68	**86.47**	**87.84**	**87.15**
JNLPBA	23.51	49.53	31.89	15.43	33.74	21.18	**26.97**	**48.34**	**34.62**	2.50	15.32	4.30	**72.24**	**83.56**	**77.49**
NCBI-disease	24.76	51.25	33.39	25.10	41.04	31.15	**39.33**	**54.69**	**45.75**	2.56	21.67	4.58	**88.22**	**91.25**	**89.71**
linnaeus	2.87	24.84	5.14	3.81	20.80	6.44	**8.30**	**42.92**	**13.91**	0.73	24.21	1.42	**90.77**	**85.83**	**88.24**
s800	9.38	45.89	15.57	10.80	39.50	16.96	**15.74**	**51.11**	**24.07**	0.99	17.21	1.87	**72.80**	**75.36**	**74.06**

Table 10
Performance on various summarization datasets. Here, 'R-1', 'R-2', 'R-L' and 'B-S' denote 'ROUGE-1', 'ROUGE-2', 'ROUGE-L', and 'BERTScore', respectively. State-of-the-art (SOTA) results are taken from the BioBART [4] model. Also, LLaMA-2 refers to its 13b version, similar to other tasks.

Model	Dataset																							
	iCliniq				HealthCareMagic				MeQSum				MEDIQA-QS				MEDIQA-MAS				MEDIQA-ANS			
	R-1	R-2	R-L	B-S	R-1	R-2	R-L	B-S	R-1	R-2	R-L	B-S	R-1	R-2	R-L	B-S	R-1	R-2	R-L	B-S	R-1	R-2	R-L	B-S
GPT-3.5	**30.5**	**12.8**	**25.4**	**89.3**	**28.1**	**9.8**	**24.0**	**88.9**	30.0	12.3	26.2	89.0	30.6	11.6	26.7	89.0	**38.9**	**14.6**	**22.1**	**87.9**	**28.7**	**10.4**	**24.4**	**89.0**
PaLM-2	21.9	10.2	18.6	87.0	25.9	9.8	22.0	88.3	31.5	14.0	27.7	89.7	29.7	11.5	26.0	90.0	15.3	8.6	13.5	85.2	25.4	12.1	18.9	85.4
Claude-2	28.8	11.0	23.7	89.0	24.4	7.4	20.3	88.2	**31.7**	**13.6**	**27.9**	**89.9**	**32.0**	**13.5**	**27.7**	90.2	13.4	6.2	11.1	85.6	28.6	8.7	17.6	85.9
LLaMA-2	20.0	7.2	15.2	85.8	16.7	5.1	12.9	85.3	21.2	7.3	17.1	85.5	23.3	8.6	17.7	86.2	13.7	11.2	13.2	86.6	28.0	9.6	17.4	85.3
SOTA	**61.1**	**48.5**	**59.4**	**94.1**	**46.7**	**26.1**	**44.2**	**91.9**	**55.6**	**38.1**	**53.2**	**93.3**	**32.0**	12.4	**29.7**	**90.3**	32.9	11.3	29.3	86.1	21.6	9.3	19.2	85.7

Table 11
Performance on the Biomedical Lay Summarization task. State-of-the-Art results are from Sim et al. [84].

Model	Dataset							
	eLife				PLOS			
	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
GPT-3.5	33.88	8.64	17.15	**84.49**	**41.11**	**11.41**	**21.74**	**86.11**
PaLM-2	21.55	3.92	12.14	81.03	29.61	7.10	16.40	83.02
Claude-2	**39.20**	**9.31**	**18.34**	84.30	39.05	9.28	19.52	85.03
LLaMA-2-13b	38.53	8.69	18.10	83.18	38.58	11.15	20.14	84.69
State-of-the-Art	**49.50**	**14.60**	**46.90**	**85.50**	**50.20**	**19.00**	**46.20**	**86.50**

Table 12
Performance on Readability Controlled Summarization in the PLOS dataset. State-of-the-Art results are from Chen et al. [85].

Model	Summarization type							
	Abstract				Lay Summarization			
	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
GPT-3.5	39.65	11.01	20.76	**85.64**	**39.13**	**9.57**	**20.00**	**85.63**
PaLM-2	25.09	5.37	14.20	82.53	30.70	7.02	16.39	83.31
Claude-2	**42.25**	**13.05**	**21.53**	85.46	36.16	7.82	17.68	84.47
LLaMA-2-13b	41.78	13.01	21.37	84.63	36.33	9.53	18.89	84.18
State-of-the-Art	**46.97**	**15.57**	**42.87**	85.48	**45.67**	**13.38**	**41.59**	85.57

training data and GPT-3.5 and other LLMs usually achieve comparable or even better performance in these datasets compared to the BioBART model fine-tuned on other related datasets [4]. This further confirms that zero-shot LLMs are more useful than domain-specific fine-tuned models in biomedical datasets that lack large training data.

We also present our findings on the biomedical lay summarization task in Table 11 and readability controlled summarization task in Table 12.

For the biomedical lay summarization task, we combine both abstract and article together and give as input to the models till the concatenated text reaches the maximum context length. For this task, we compare the performance of the LLMs in eLife and PLOS datasets. Based on the ROUGE scores, the Claude-2 model is found to be the best performing LLM in the eLife dataset with GPT-3.5 being the best-performing one in the PLOS dataset. However, none of the LLMs could outperform the current state-of-the-art in these datasets. While the performance of the LLMs is quite low in terms of ROUGE, they achieve much higher scores in terms of BERTScore, which is comparable to the state-of-the-art result. This shows a great discrepancy between the lexical matching based traditional ROUGE scoring and the contextual similarity-based BERTScore metric.

The readability-controlled summarization task contains two sub-tasks: (i) abstract writing, and (ii) lay summary writing. Contrary to the previous task (i.e., biomedical lay summarization task), this time we only give an article as input without the abstract, as required by the task. We find that in writing the abstract of the given article, the Claude-2 model performs the best in terms of all ROUGE scores. However, in terms of BERTScore, GPT-3.5 slightly performs better than Claude-2. Interestingly, we find that in terms of the BERTScore, the GPT-3.5 model even outperforms the ROUGE-based SOTA models in both datasets. This further establishes the limitation of using ROUGE as a metric to evaluate LLMs for summarization [75].

Since the whole document cannot be given as input at once to these LLMs except Claude-2, we also investigate the performance using the

Table 13
Performance of different LLMs on Biomedical Lay Summarization datasets based on various input lengths.

Model	Length	Dataset							
		eLife				PLOS			
		ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
GPT-3.5	2000	33.88	8.64	17.15	**84.49**	41.11	11.41	21.74	86.11
GPT-3.5	5000	33.62	8.77	17.21	84.45	41.41	11.65	21.89	86.17
GPT-3.5	10 000	33.39	8.60	17.16	84.35	**41.59**	**11.94**	**22.11**	**86.25**
PaLM-2	2000	21.55	3.92	12.14	81.03	29.61	7.10	16.40	83.02
PaLM-2	5000	15.13	2.54	8.71	79.27	25.00	5.78	13.89	82.10
Claude-2	2000	39.20	9.31	18.34	84.30	39.05	9.28	19.52	85.03
Claude-2	5000	**39.43**	**9.42**	**18.38**	84.20	38.79	9.09	19.26	84.92
Claude-2	FULL	38.97	9.09	18.05	83.95	39.16	9.31	19.30	84.85

Table 14
Performance of different LLMs on Readability Controlled Summarization in the PLOS dataset based on various input lengths.

Model	Length	Summarization type							
		Abstract				Lay summarization			
		ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
GPT-3.5	2000	39.65	11.01	20.76	85.64	39.13	9.57	20.00	85.53
GPT-3.5	5000	40.94	11.83	21.40	85.90	40.07	10.27	20.66	85.81
GPT-3.5	10 000	40.99	11.89	21.44	**85.91**	**40.29**	**10.42**	**20.71**	**85.86**
PaLM-2	2000	25.09	5.37	14.20	82.53	30.70	7.02	16.39	83.31
PaLM-2	5000	21.98	4.63	12.38	81.55	25.05	5.43	13.81	82.03
Claude-2	2000	42.25	13.05	21.53	85.46	36.16	7.82	17.68	84.47
Claude-2	5000	43.27	13.60	22.29	85.67	37.97	8.58	18.56	84.66
Claude-2	FULL	**43.89**	**13.88**	**22.49**	85.72	38.97	9.09	18.05	83.95

following input context lengths (in terms of number of words); PaLM-2: 2000 and 5000, GPT-3.5: 2000, 5000, and 10000, and Claude-2: 2000, 5000, and full input document. Since LLaMA-2 has a maximum context length of 4000 tokens (approximately 3000 words[15]), we exclude LLaMA-2 from this study. The results for both tasks, biomedical lay summarization, and readability controlled summarization, can be found in Table 13 and Table 14, respectively. Our experiments reveal that increasing the context length decreases the performance of PaLM-2 in both tasks across all datasets. Moreover, increasing the context length also does not help GPT-3.5 or Claude-2 to gain any substantial performance gain. This can be explained based on the work of Liu et al. [88], where they find that LLMs tend to lose contextual information with the increase in sequence length, and especially they perform poorly in scenarios when they are required to generate responses based on utilizing the information that appears in the middle of the context.

The experimental results in these article summarization datasets demonstrate that using the context length of 2000 is good enough in terms of ROUGE and BERTScore metrics for both abstract and lay summarization. This context length should also be very helpful in terms of usage cost as well as time efficiency in comparison to using longer contexts [89]

Further performance analysis demonstrates that based on the paired t-test with $p \leq .05$, the performance difference in terms of the ROUGE score between all the LLMs and the current fine-tuned SOTA models in the summarization datasets **is statistically significant**, which also happens in terms of BERTScore for all LLMs except GPT-3.5.

5.4. Analysis

In this section, we conduct further analysis on the performance of LLMs based on (i) variations in prompts, (ii) few-shot learning, and (iii) fine-tuning, alongside analyzing the performance of LLMs based on the (iv) possibility of data contamination. Below, the findings based on this analysis are demonstrated.

5.4.1. Effects of prompt variation

The effects of prompt tuning in the HoC dataset have been investigated by evaluating the performance of GPT-3.5 based on the following prompt variations:

i. Prompting with explicitly defining the 10 HoC classes achieves an F1 score of 59.26 (see Row 1 in Table 15).
ii. Prompting without mentioning the name of any HoC classes, drops F1 to 38.20 (see Row 2 in Table 15).
iii. Prompting with the name of each HoC class is given without providing the definition of each class, drops the F1 score to 46.93 (see Row 3 in Table 15).

Thus, our findings demonstrate that more descriptive prompts yield better results.

5.4.2. Effects of few-shot learning

In the previous analysis, it has been found that variations in prompts, especially the utilization of more descriptive prompts could significantly impact the performance of LLMs in zero-shot scenarios. While the main focus of this work was to conduct zero-shot experiments using LLMs to address the lack of large annotated datasets in the biomedical domain, this section demonstrates the effect of the utilization of few-shot examples in the prompts. Since few-shot learning also leads to an increase in the context length, which is a problem for LLMs that have limited context length, in this paper, the Claude-2 model is selected for the few-shot experiments since it can consider significantly much longer contexts (100k tokens) than other LLMs. Thus, using Claude-2 as the LLM for the few-shot learning experiments also helped us to address the context length issue. In the prompt, the few-shot examples are first included, followed by the task descriptions, as demonstrated in Section 4.1. The results from the few-shot experiments across all datasets are shown in Table 16.

[15] https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them

Table 15
Effects of Prompt Variations in GPT-3.5 for the Document Classification Task in the HoC dataset.

#	Prompt	F1
1.	The 10 hallmarks of cancer taxonomy with their definitions are given below: (i) Sustaining proliferative signaling: Cancer cells can initiate and maintain continuous cell division by producing their own growth factors or by altering the sensitivity of receptors to growth factors. (ii) Evading growth suppressors: Cancer cells can bypass the normal cellular mechanisms that limit cell division and growth, such as the inactivation of tumor suppressor genes and/or insensitivity to antigrowth signals. (iii) Resisting cell death: Cancer cells develop resistance to apoptosis, the programmed cell death process, which allows them to survive and continue dividing. (iv) Enabling replicative immortality: Cancer cells can extend their ability to divide indefinitely by maintaining the length of telomeres, the protective end caps on chromosomes. (v) Inducing angiogenesis: Cancer cells stimulate the growth of new blood vessels, providing the necessary nutrients and oxygen to support their rapid growth. (vi) Activating invasion and metastasis: Cancer cells can invade surrounding tissues and migrate to distant sites in the body, forming secondary tumors called metastases. (vii) Deregulating cellular energetic metabolism: Cancer cells rewire their metabolism to support rapid cell division and growth, often relying more on glycolysis even in the presence of oxygen (a phenomenon known as the Warburg effect). (viii) Avoiding immune destruction: Cancer cells can avoid detection and elimination by the immune system through various mechanisms, such as downregulating cell surface markers or producing immunosuppressive signals. (ix) Tumor promoting inflammation: Chronic inflammation can promote the development and progression of cancer by supplying growth factors, survival signals, and other molecules that facilitate cancer cell proliferation and survival. (x) Genome instability and mutation: Cancer cells exhibit increased genomic instability, leading to a higher mutation rate, which in turn drives the initiation and progression of cancer. Classify the following sentence in one of the above 10 hallmarks of cancer taxonomy. If cannot be classified, answer as “empty”: [SENTENCE]	59.26
2.	Is it possible to classify the following sentence in one of the 10 categories in the Hallmarks of Cancer taxonomy? If possible, write down the class. [SENTENCE]	38.20
3.	Classify the sentence given below in one of the 10 categories (i. activating invasion and metastasis, ii. tumor promoting inflammation, iii. inducing angiogenesis, iv. evading growth suppressors, v. resisting cell death,vi. cellular energetics, vii. genomic instability and mutation, viii. sustaining proliferative signaling, ix. avoiding immune destruction, x. enabling replicative immortality) in the Hallmarks of Cancer taxonomy? If cannot be classified, answer as “empty”. [SENTENCE]	46.93

Though few-shot learning usually leads to improvements in performance, in many tasks, few-shot learning is also found to be ineffective. For instance, Ye et al. [90] demonstrated that in many language processing tasks, few-shot learning using LLMs achieves much poorer results in comparison to zero-shot learning. In our experiments, we also find that while few-shot learning is more effective than zero-shot in some tasks (e.g., better in terms of F1 in KD-DTI (1-shot) and BC5CDR (3-shot) for relation extraction,[16] in terms of Accuracy in MediQA-2019 (1-shot) and PubMedQA (3-shot) for question answering, as well as in some summarization datasets), the opposite happens in other tasks as well (e.g., NER, Entity Linking, etc.). Therefore our findings are consistent with Ye et al. [90] to reveal that increasing few-shot examples from 0-shot to 1 or 3-shot does not necessarily improve the performance.

To further improve performance with few-shot, the task examples in few-shot prompts are required to be of high quality to ensure better performance while avoiding possible prediction biases towards the task examples. Thus, future work may investigate how to construct better examples for few-shot experiments with LLMs in the biomedical domain.

5.4.3. Effects of fine-tuning

The few-shot learning experiment demonstrates that adding few-shot examples to the prompt does not lead to any performance gain in most biomedical tasks. Thus, in this section, we investigate whether the fine-tuning of LLMs could lead to performance gain. Since the main motivation of this paper is to investigate how LLMs could be used to address the lack of annotated datasets problem in the biomedical domain, only the datasets that have smaller training sets have been used for the fine-tuning experiment. This makes the fine-tuning experiment to be also consistent with the motivation of this paper which is to investigate the capability of LLMs in zero-shot scenarios in the biomedical domain to address the lack of large annotated dataset issue. For this reason, the PubMedQA dataset for question-answering (only 450 training samples), the MeQSum dataset (500 training samples) for summarization, the DDI (500 training samples), and the BC5CDR (664 training samples) datasets for relation extraction have been used for LLM fine-tuning. Nonetheless, many closed-source LLMs (e.g., PaLM-2, Claude-2) do not support fine-tuning, whereas fine-tuning GPT-3.5 significantly increases the cost during inference.[17] Thus, the fine-tuning experiment is conducted with a comparatively smaller open-source LLM: the LLaMA-2-7B-Chat[18] model and run for 3 epochs with the learning rate $2e-5$. These hyperparameters are selected since they lead to the best performance in the validation set. The results of the fine-tuning experiment are shown in Table 17. From Table 17, it is quite evident that fine-tuning is more useful than few-shot learning. In general, fine-tuning outperforms all the zero-shot and few-shot LLMs (except GPT-3.5 in the DDI dataset in terms of Recall and F1, even though the fine-tuned version achieves significantly better precision scores). Meanwhile, in the summarization dataset, the fine-tuned LLaMA-2-7B set a new state-of-the-art result. Moreover, it achieves almost similar performance in comparison to the state-of-the-art in the PubMedQA dataset for the question answering task (even though LLaMA-2-7B was only trained on 500 samples, the current state-of-the-art BioGPT [5] model was trained on 270 K samples).

[16] Few-shot learning leads to a decrease in performance in terms of Recall in comparison to zero-shot learning in all relation extraction datasets.

[17] https://openai.com/blog/gpt-3-5-turbo-fine-tuning-and-api-updates

[18] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

Table 16
Experimental Results for Few-Shot Learning. Here, 'Readability-Controlled', 'ROUGE', and 'BERTScore' are denoted by 'RC', 'R', and 'B-S', respectively.'.

Dataset	Claude-2 (0-Shot)			Claude-2 (1-Shot)			Claude-2 (3-Shot)			SOTA		
NER	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
BC2GM	31.95	55.10	40.45	29.88	51.89	37.92	29.76	47.19	36.50	**84.32**	**85.12**	**84.72**
BC4CHEMD	26.37	52.83	35.18	22.28	52.41	31.27	26.87	51.12	35.23	**92.80**	**91.92**	**92.36**
BC5CDR-chem	49.99	69.23	58.05	46.27	59.07	51.89	49.27	65.61	56.28	**93.68**	**93.26**	**93.47**
BC5CDR-disease	47.06	53.62	50.13	44.65	52.71	48.35	43.77	51.27	47.22	**86.47**	**87.84**	**87.15**
JNLPBA	26.97	48.34	34.62	26.63	46.29	33.81	27.38	44.11	33.79	**72.24**	**83.56**	**77.49**
NCBI-disease	39.33	54.69	45.75	37.28	55.42	44.57	35.69	49.48	41.47	**88.22**	**91.25**	**89.71**
linnaeus	8.30	42.92	13.91	8.31	33.22	13.29	14.43	40.13	21.23	**90.77**	**85.83**	**88.24**
s800	15.74	51.11	24.07	19.54	49.54	28.02	15.45	47.59	23.32	**72.80**	**75.36**	**74.06**
Relation Extraction	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
BC5CDR	44.04	**67.73**	53.37	**66.95**	40.45	50.18	62.17	53.34	**57.42**	49.52	43.25	46.17
KD-DTI	17.99	**72.73**	28.84	39.43	55.32	**46.04**	36.80	13.93	20.21	**40.00**	39.72	38.42
DDI	39.27	**46.60**	**42.62**	30.69	28.80	29.72	33.89	24.27	28.28	**41.70**	44.75	40.76
Entity Linking	Recall@1			Recall@1			Recall@1			Recall@1		
BC5CDR	78.01			47.91			55.68			**93.26**		
Cometa	53.29			55.59			56.99			**81.77**		
NCBI	70.21			49.17			47.60			**89.90**		
Question Answering	Accuracy			Accuracy			Accuracy			Accuracy		
PubMedQA	57.20			52.23			62.80			**78.20**		
MediQA-2019	65.13			68.65			63.32			**79.49**		
Text Classification	F1			F1			F1			F1		
HoC	34.93			38.99			43.78			**85.12**		
LitCovid	7.60			4.01			6.27			**86.20**		
Summarization	R-1/R-2/R-L/B-S			R-1/R-2/R-L/B-S			R-1/R-2/R-L/B-S			R-1/R-2/R-L/B-S		
iCliniq	28.8/11.0/23.7/89.0			30.9/12.4/25.9/88.9			29.8/11.4/24.2/88.8			**61.1/48.5/59.4/94.1**		
HealthCareMagic	24.4/7.4/20.3/88.2			24.9/7.2/20.4/87.7			24.9/7.9/20.6/87.9			**46.7/26.1/44.2/91.9**		
MeQSum	31.7/13.6/27.9/89.9			26.8/10.6/22.4/87.7			29.1/11.7/24.8/88.2			**55.6/38.1/53.2/93.3**		
MEDIQA-QS	32.0/**13.5**/27.7/90.2			26.8/11.0/21.8/88.1			27.7/11.0/22.02/88.2			**32.0**/12.4/**29.7/90.3**		
MEDIQA-MAS	13.4/6.2/11.1/85.6			**36.5/11.4/20.3/86.7**			36.3/11.4/20.3/86.7			32.9/11.3/29.3/86.1		
MEDIQA-ANS	28.6/8.7/17.6/85.9			30.9/10.8/19.6/86.3			**31.5/11.8/20.7/86.5**			21.6/9.3/19.2/85.7		
eLife (Lay Summ)	39.2/9.3/18.3/84.3			39.3/8.9/17.9/84.1			37.6/8.5/17.5/84.1			**49.5/14.6/46.9/85.5**		
PLOS (Lay Summ)	39.1/9.3/19.5/85.0			38.7/8.8/18.8/84.8			38.8/8.83/18.9/84.9			**50.2/19.0/46.2/86.5**		
PLOS (RC: Abstract)	42.3/13.1/21.5/85.5			42.4/12.8/21.5/85.4			42.7/12.7/21.5/85.5			**47.0/15.6/42.9/85.5**		
PLOS (RC: Lay Summ)	36.2/7.8/17.7/84.5			38.0/8.2/18.3/84.6			37.1/7.7/17.8/84.5			**45.7/13.4/41.6/85.6**		

Table 17
Experimental Results for Fine-Tuning. Here, 'ROUGE' and 'BERTScore' are denoted by 'R' and 'B-S', respectively.

Model	Relation Extraction Task						QA Task	Summarization Task			
	BC5CDR			DDI			PubMedQA	MeQSum			
	Precision	Recall	F1	Precision	Recall	F1	Accuracy	R-1	R-1	R-L	B-S
GPT-3.5 (0-Shot)	30.62	73.85	43.29	47.11	45.77	**46.43**	54.40	30.0	12.3	26.2	89.0
PaLM-2 (0-Shot)	51.61	57.30	54.30	35.47	16.48	22.50	59.60	31.5	14.0	27.7	89.7
Claude-2 (0-Shot)	44.04	67.73	53.37	39.27	**46.60**	42.62	57.20	31.7	13.6	27.9	89.9
LLaMA-2-13b (0-Shot)	39.54	**81.66**	53.28	22.58	25.67	24.03	61.40	21.2	7.3	17.1	85.5
Claude-2 (1-Shot)	66.95	40.45	50.18	30.69	28.80	29.72	52.23	26.8	10.6	22.4	87.7
Claude-2 (3-Shot)	62.17	53.34	57.4	33.89	24.27	28.28	62.80	29.1	11.7	24.8	88.2
LLaMA-2-7b (Fine-Tuned)	**69.28**	49.86	**57.99**	**60.57**	32.15	42.00	78.00	**55.8**	**38.4**	**53.6**	**91.7**
SOTA	49.52	43.25	46.17	41.70	44.75	40.76	**78.20**	55.6	38.1	53.2	93.3

5.4.4. Data contamination detection analysis

We follow the work of Li et al. [91] to analyze the possibility of the contamination of the datasets that we study in this paper to evaluate various LLMs. For this purpose, we do the following similar[19] to their work [91].

i. **Task Example Extraction:** This contamination detection technique checks whether the task example of a particular dataset (evaluated on discriminative tasks, i.e., non-summarization) can be extracted from the LLMs that we evaluated in this paper.
ii. **Membership Inference:** This contamination detection technique checks whether the response generated by LLMs in a particular dataset (evaluated on generation tasks, i.e., summarization) is an exact match of any gold labels in that dataset.

The results of the data contamination detection analysis are shown in Table 18. From Table 18, it can be inferred that in the NER datasets, none of the LLMs could extract the task examples. This could be due

[19] We did not compare the performance of LLMs based on the chronological analysis (which was also used by Li et al. in [91]) since most of the classification datasets that have been used in this paper came before the data cut-off date of different LLMs.

Table 18
Data contamination detection analysis. Here, 'Task Example Extraction' and 'Membership Inference' are denoted by 'TEE' and 'MI', respectively; whereas 'NO' indicates that the possibility of contamination is not found, and 'YES' indicates that there is a possibility of contamination found.

Task & Dataset	GPT-3.5	PaLM-2	Claude-2	LLaMA-2-13B
NER	TEE	TEE	TEE	TEE
BC2GM (2008)	No	No	No	No
BC4CHEMD (2016)	No	No	No	No
BC5CDR-chem (2015)	No	No	No	No
BC5CDR-disease (2014)	No	No	No	No
JNLPBA (2004)	No	No	No	No
NCBI-disease (2016)	No	No	No	No
linnaeus (2010)	No	No	No	No
s800 (2013)	No	No	No	No
Relation Extraction	TEE	TEE	TEE	TEE
BC5CDR (2016)	No	No	No	No
KD-DTI (2022)	No	Yes	No	No
DDI (2013)	Yes	No	Yes	No
Entity Linking	TEE	TEE	TEE	TEE
BC5CDR	No	No	No	No
Cometa	No	No	No	No
NCBI	No	No	No	No
Question Answering	TEE	TEE	TEE	TEE
PubMedQA (2019)	No	No	Yes	Yes
MediQA-2019 (2019)	No	No	No	No
Text Classification	TEE	TEE	TEE	TEE
HoC (2016)	No	No	No	No
LitCovid (2020)	No	Yes	No	No
Summarization	MI	MI	MI	MI
iCliniq (2020)	No	Yes	Yes	No
HealthCareMagic (2020)	Yes	Yes	Yes	Yes
MeQSum (2019)	Yes	Yes	Yes	Yes
MEDIQA-QS (2021)	No	No	No	No
MEDIQA-ANS (2020)	No	Yes	No	No
MEDIQA-MAS (2021)	No	No	No	No
eLife (Lay Summ) (2023)	No	No	No	No
PLOS (Lay Summ) (2023)	No	No	No	No
PLOS (RC: Abstract) (2023)	No	No	No	No
PLOS (RC: Lay Summ) (2023)	No	No	No	No

Table 19
Results on NER and text classification based on accuracy.

Model	Dataset									
	NER								Text classification	
	BC2GM	BC4CHEMD	BC5CDR-chem	BC5CDR-disease	JNLPBA	NCBI-disease	linnaeus	s800	HoC	LitCovid
GPT-3.5	85.00	84.14	87.26	87.98	80.53	87.90	87.71	87.15	58.86	26.55
PaLM-2	88.36	87.66	92.60	91.75	82.65	89.62	93.37	91.16	62.64	43.02
Claude-2	89.92	89.86	94.03	94.55	84.59	92.32	93.72	92.41	35.80	0.87
LLaMA-2	50.78	49.78	54.99	50.79	47.59	47.66	50.46	46.43	42.89	13.39

to the fact that in our experiments, the LLMs were asked to determine the NER tag for each token based on the 'BIO' format. Meanwhile, the LLMs could potentially be pre-trained differently for the NER task. In our analysis, we also find that while LLMs could explain the NER tasks, they cannot generate the task examples for each dataset in the expected 'BIO format'. The experimental results demonstrate that the possible absence of the task examples in the pre-training data could probably be the reason behind LLMs performing very poorly in all NER datasets. A similar trend is also observed in the Entity Linking datasets where no possibility of data contamination is found based on the task extraction analysis technique.

However, in Relation Extraction, task examples could be extracted in the KD-DTI and the DDI datasets (while the task example extraction approach did not lead to the possibility of data contamination in BC5CDR). In the case of the KD-DTI dataset, the best-performing PaLM-2 model could extract task examples, whereas in the DDI dataset, two of the better-performing LLMs, GPT-3.5 and Claude-2, could also extract task examples. This may indicate that the possible presence of task examples in the LLM training data may be responsible for the improved performance of some LLMs in respective datasets.

In terms of the question answering and the text classification datasets, the task example extraction techniques show no possibility of data contamination in MediQA-2019 and HoC datasets. This is quite surprising for GPT-3.5 in the MediQA-2019 dataset since it achieves performance comparable to the state-of-the-art. While for HoC, it is

Table 20
Some examples of PaLM-2 generated responses in Relation Extraction and Text Classification.

Dataset	Type	Prompt	PaLM-2 Response	Gold Label
BC5CDR	Chemical-Disease Relation Extraction	Identify each pair of drugs and the drug-induced side-effects (e.g., diseases) in the following passage: Famotidine-associated delirium. A series of six cases. Famotidine is a histamine H2-receptor antagonist used in inpatient settings for prevention of stress ulcers and is showing increasing popularity because of its low cost. Although all of the currently available H2-receptor antagonists have shown the propensity to cause delirium, only two previously reported cases have been associated with famotidine. The authors report on six cases of famotidine-associated delirium in hospitalized patients who cleared completely upon removal of famotidine. The pharmacokinetics of famotidine are reviewed, with no change in its metabolism in the elderly population seen. The implications of using famotidine in elderly persons are discussed.	The drug is famotidine. The drug induced disease is delirium. The side effect is delirium.	chemical: famotidine disease: delirium
KD-DTI	Drug-Target Relation Extraction	Identify the drug-target interactions in the following passage (along with the interaction type among the following: ’inhibitor’, ’agonist’, ’modulator’, ’activator’, ’blocker’, ’inducer’, ’antagonist’, ’cleavage’, ’disruption’, ’intercalation’, ’inactivator’, ’bind’, ’binder’, ’partial agonist’, ’cofactor’, ’substrate’, ’ligand’, ’chelator’, ’downregulator’, ’other’, ’antibody’, ’other/unknown’): Binding of valsartan to mammalian angiotensin AT1 receptors. The binding characteristics of the angiotensin AT1 receptor antagonist valsartan were investigated in different animal species and tissues. Using [125I](Sar1,Ile8) angiotensin II as radioligand, affinity constants were determined in liver and adrenal rat and marmoset, human adrenal and in rat aortic smooth muscle cells. In all tissues tested, valsartan had a greater affinity for the AT1 receptor than losartan (on average 5-fold). The affinities of both antagonists were up to 30 times weaker in the dog tissues [3H]Valsartan bound with high affinity (Kd 1.44 nmol/l) to the rat aortic smooth muscle cell AT1 receptor. Binding was saturable and reversible. Non-specific binding was low (10%). Reports that [3H]losartan binds to a non-angiotensin II binding site in rat liver and in other tissues could be confirmed. [3H]Valsartan on the other hand bound only to the AT1 receptor. Using a competition binding assay with [3H]losartan on rat liver membranes it could be shown that valsartan can bind to the ’losartan binding site’, but at a 10,000-fold less affinity than for the AT1 receptor. Valsartan is therefore a highly specific and selective antagonist of the AT1 receptor. Due to its high affinity and low non-specific binding it is a suitable radioactive antagonist for the study of the distribution and function of the angiotensin AT1 receptor.	valsartan, antagonist, angiotensin AT1 receptor	drug: Valsartan target: Type-1 angiotensin II receptor interaction: antagonist
DDI	Drug–Drug Relation Extraction	Identify the pairs of drug–drug interactions in the passage given below based on one of the following interaction types: (i) mechanism: this type is used to identify drug–drug interactions that are described by their pharmacokinetic mechanism. (ii) effect: this type is used to identify drug–drug interactions describing an effect. (iii) advice: this type is used when a recommendation or advice regarding a drug–drug interaction is given. (iv) int: this type is used when a drug–drug interaction appears in the text without providing any additional information. Interactions between Leukine and other drugs have not been fully evaluated. Drugs which may potentiate the myeloproliferative effects of Leukine, such as lithium and corticosteroids, should be used with caution.	Leukine, lithium: effect	drug: Leukine target: lithium interaction: effect drug: Leukine target: corticosteroids interaction: effect
HoC	Text Classification	The 10 hallmarks of cancer taxonomy with their definitions are given below: (i) Sustaining proliferative signaling: Cancer cells can initiate and maintain continuous cell division by producing their own growth factors or by altering the sensitivity of receptors to growth factors. (ii) Evading growth suppressors: Cancer cells can bypass the normal cellular mechanisms that limit cell division and growth, such as the inactivation of tumor suppressor genes and/or insensitivity to antigrowth signals. (iii) Resisting cell death: Cancer cells develop resistance to apoptosis, the programmed cell death process, which allows them to survive and continue dividing. (iv) Enabling replicative immortality: Cancer cells can extend their ability to divide indefinitely by maintaining the length of telomeres, the protective end caps on chromosomes. (v) Inducing angiogenesis: Cancer cells stimulate the growth of new blood vessels, providing the necessary nutrients and oxygen to support their rapid growth. (vi) Activating invasion and metastasis: Cancer cells can invade surrounding tissues and migrate to distant sites in the body, forming secondary tumors called metastases. (vii) Deregulating cellular energetic metabolism: Cancer cells rewire their metabolism to support rapid cell division and growth, often relying more on glycolysis even in the presence of oxygen (a phenomenon known as the Warburg effect). (viii) Avoiding immune destruction: Cancer cells can avoid detection and elimination by the immune system through various mechanisms, such as downregulating cell surface markers or producing immunosuppressive signals. (ix) Tumor promoting inflammation: Chronic inflammation can promote the development and progression of cancer by supplying growth factors, survival signals, and other molecules that facilitate cancer cell proliferation and survival. (x) Genome instability and mutation: Cancer cells exhibit increased genomic instability, leading to a higher mutation rate, which in turn drives the initiation and progression of cancer. Classify the following sentence in one of the above 10 hallmarks of cancer taxonomy. If cannot be classified, answer as “empty”: In addition, activated phosphatidylinositol 3-kinase (PI3-K) and phosphorylated Akt levels were also elevated in the U87MG.DeltaEGFR tumors.	sustaining proliferative signaling.	sustaining proliferative signaling.

expected since all LLMs perform much poorer than the state-of-the-art. For the other question-answering and text classification datasets, LLaMA-2 and Claude-2 show the possibility of data contamination in the PubMedQA dataset. This may provide some explanations on why smaller LLaMA-2-13b outperforms other much larger LLMs in this dataset. In the LitCovid dataset, we only find that the PaLM-2 model has the possibility of data contamination (it also achieves the best result in comparison to other LLMs in this dataset).

In the summarization datasets, the contamination detection analysis is conducted based on the membership inference technique which demonstrates that PaLM-2 is more likely to generate some responses similar to the gold reference summaries, as it shows the possibility of membership inference-based contamination in the highest number of datasets (4 out of the 10 summarization datasets). We also find that the HealthcareMagic and the MeQSum datasets are reported as contaminated based on membership inference for all four LLMs. However, in none of these datasets, LLMs could beat the state-of-the-art models

Table 21
Some examples of Claude-2 generated responses in Entity Linking, Named Entity Recognition, and Question Answering.

Dataset	Type	Prompt	Claude-2 Response	Gold Label
Cometa	Entity Linking	“The season seemed to last forever, and even with START allegra END d, i was still feeling pretty crappy.” In the biomedical text given above, what does the entity between the START and END token refer to (answer without any explanation)?	allegra	fexofenadine
BC2GM	Named Entity Recognition	Below, we provide a biomedical text: “Previous analysis of this motif in the lactose permease (A .” You need to identify the GENE/PROTEIN type named entities in the above text. To identify the named entities, please tag each token of the given text in the ’BIO’ format as either: ’B’ or ’I’ or ’O’. The BIO format stands for Beginning, Inside, Outside. It provides a way to label individual tokens in a given text to indicate whether they are part of a named entity. In the BIO format, each token in a text is labeled with a tag that represents its role in a named entity. For our case, there are three possible tags: B: it indicates that the token is the beginning of the GENE/PROTEIN type named entity (i.e., the first token of a GENE/PROTEIN type named entity). I: it indicates the token is inside a GENE/PROTEIN type named entity (i.e., any token other than the first token of a GENE/PROTEIN type named entity). O: it indicates that the token is outside any named entity. In other words, it is not part of any named entity. Below, each token of the biomedical text is provided (separated by new line). Now please assign the correct tag to each token. Return your result for each token in a newline in the following format -> token: assigned tag. Previous : <ASSIGNED TAG> analysis : <ASSIGNED TAG> of : <ASSIGNED TAG> this : <ASSIGNED TAG> motif : <ASSIGNED TAG> in : <ASSIGNED TAG> the : <ASSIGNED TAG> lactose : <ASSIGNED TAG> permease : <ASSIGNED TAG> (: <ASSIGNED TAG> A : <ASSIGNED TAG> . : <ASSIGNED TAG>	Previous : O analysis : O of : O this : O motif : O in : O the : O lactose : B permease : I (: O A : B . : O	O O O O O O O B I O O O
PubMedQA	Question Answering	For the question, the reference context, and the answer given below, is it possible to infer the answer for that question from the reference context? Only reply as either Yes or No or Maybe. Question: Do some U.S. states have higher/lower injury mortality rates than others? Reference context: This article examines the hypothesis that the six U.S. states with the highest rates of road traffic deaths (group 1 states) also had above-average rates of other forms of injury such as falling, poisoning, drowning, fire, suffocation, homicide, and suicide, and also for the retail trade and construction industries. The converse, second hypothesis, for the six states with the lowest rates of road traffic deaths (group 2 states) is also examined. Data for these 12 states for the period 1983 to 1995 included nine categories of unintentional and four categories of intentional injury. Seventy-four percent of the group 1 states conformed to the first hypothesis, and 85% of the group 2 states conformed to the second hypothesis. Answer: Group 1 states are likely to exhibit above-average rates for most other categories of injury death, whereas group 2 states are even more likely to exhibit below-average rates for most other categories of injury death.	Yes.	Yes.

(with the results being much lower in comparison to the reported state-of-the-art results). It should also be pointed out that the membership inference shows no possibility of contamination in datasets that are released in 2023.

6. Conclusions and future work

In this paper, we evaluate LLMs in six benchmark biomedical tasks across 26 datasets. We observe that in datasets that have large training data, zero-shot LLMs usually fail to outperform the fine-tuned state-of-the-art models (e.g., BioBERT, BioGPT, BioBART, etc.). However, they consistently outperform the fine-tuned baselines on tasks where the state-of- the-art results were achieved based on fine-tuning only on smaller training sets. While the LLMs that are studied in this paper are massive language models with a billion of parameters, they are trained on diverse domains and so when evaluating their zero-shot capabilities, they usually fail to outperform various state-of-the-art biomedical task specific fine-tuned models. However, fine-tuning these LLMs even on smaller training sets significantly improves their performance. Thus, it could be useful to train biomedical domain-specific LLMs on biomedical corpora to achieve better performance in tasks related to the biological and the medicine domain. Moreover, our findings demonstrate that the performance of these LLMs may vary across different datasets and tasks, as we did not observe a single LLM outperforming others across all datasets and tasks. Thus, our evaluation in this paper could give a good direction for future research as well as real-world usage while utilizing these LLMs to build task-specific biomedical systems. We also demonstrate that LLMs are sensitive to prompts, as variations in prompts led to a noticeable difference in results. Thus, we believe that our evaluation will help future research while constructing the prompts for LLMs for various tasks.

In the future, we will extend our work to investigate the performance of LLMs on more biomedical tasks [14], such as medical code assignment [92], drug design [93], healthcare [26], protein sequence [94], as well as on low-resource languages [95] and problems in information retrieval that require open-domain knowledge [96–98]. We will also explore the ethical implications (e.g., privacy concerns [99]) of using LLMs in the biomedical domain. Moreover, we will extend our work to study the multi-modal LLMs [100–104] in the biomedical image processing tasks alongside also studying whether fine-tuning smaller open-source LLMs could outperform existing fine-tuned state-of-the-art models in the biomedical domain.

CRediT authorship contribution statement

Israt Jahan: Writing – review & editing, Writing – original draft, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Md Tahmid Rahman Laskar:** Writing – review & editing, Writing – original draft, Methodology, Data curation, Conceptualization. **Chun Peng:** Writing – review & editing, Writing – original draft, Validation, Supervision, Investigation. **Jimmy Xiangji Huang:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration.

Table 22
Some examples of GPT-3.5 generated responses in different Biomedical Text Summarization tasks (Dialogue Summarization, Question Summarization, and Answer Summarization).

Dataset	Type	Prompt	GPT-3.5 Response	Gold Summary
iCliniq	Dialogue Summarization	Write a very short and concise one line summary of the following dialogue as an informal question in a healthcare forum: Patient: Hello doctor, I am depressed. What can I do? My parents are not supporting for a surgery which is important for my self-confidence and self-esteem. My father is not taking care of my health or well-being. Please help. Doctor: Hello. Any related incident or cause for having depression is unclear in this query. Details of surgery and symptoms which are not mentioned are appreciated to know the complete picture. Depressive feelings are generally linked with some loss, self-related issues, stress, or environmental or biological factors. It can be managed by: 1. Medicines. 2. Psychotherapy to identify the causing agents and try resolving the issue. 3. Modifying thoughts related to concerns. 4. Physical exercises like walking, jogging, and workout. 5. Diet changes — Reduce the intake of too much salt and junk food and reduce sugar intake. Consume healthy foods and fruits having serotonin and omega-3 can reduce depression.	What are some ways to manage depression without surgery and what details should I provide to my doctor?	I am depressed, and my parents are not supporting me to overcome it. Please help
HealthCareMagic	Dialogue Summarization	Write a very short and concise one line summary of the following dialogue as a question in a healthcare forum: Patient: Dear Doctor,I have a 1.5 years old sick daughter, I am looking for the written file to understand some clinical numbers of a normal young child(1–2 years old):1: what is the diameter of the hepatic portal vein? 2: what is the frequency of the hepatic portal flow?I will appreciate if you could help me with these information. Kelvin Lu Doctor: 4–8 mm is range for diameter for hepatic portal vein for 1.5 year of children, there is no thing like frequency of hepatic portal flow. please make it clear	What is the diameter of the hepatic portal vein in a 1–2 year old child, and what is the frequency of the hepatic portal flow?	What should be the diameter of hepatic portal vein of a 1.5 years old?
MeQSum	Question Summarization	Rewrite the following question in a short and concise form: Question: Williams' syndrome . I would like to have my daughter tested for William's syndrome. Could you please tell me where I would go or who does it in my area? Thank you!!	Where can I get my daughter tested for Williams' syndrome in my area?	Where can I get genetic testing for william's syndrome?
MEDIQA-QS	Question Summarization	Rewrite the following question in a short and concise form: Question: 23 surgeries and counting......lower lip birthmark, have tried all options out the there and guess what still have it, continues to grow back.....any suggestions? Is there a cure coming in the next few years hopefully?	Any cure for a growing lower lip birthmark after 23 surgeries?	How can i get rid of a lower lip birthmark permanently?
MEDIQA-ANS	Answer Summarization	Write a very short and concise summary of the following article based on the question given below: Question: Bad Breath I have very bad breath and at times it can make myself and others sick. I need some advice as to what I need to do. Article: Bad Breath Halitosis Summary There are many reasons why you might have bad breath. You can get it if you do not brush and floss regularly. Bacteria that build up in mouth and between teeth produce the bad odor. Other problems in mouth, such as gum disease, dry mouth or cavities, may also cause it. Sinusitis or problems with nose may be to blame. You can also have bad breath if you eat some foods, like raw onions, garlic or cabbage. And of course smoking causes its own bad aroma. Some diseases and medicines are associated with a specific breath odor. Having good dental habits, like brushing and flossing regularly, help fight bad breath. Mouthwashes, mints or chewing gum may make you breath fresher. If you have an underlying disorder, treating it may help eliminate the breath odor.	The article discusses the causes and considerations of bad breath, which is usually related to poor dental hygiene. Other disorders and diseases can also produce distinct breath odors, such as fruity breath being a sign of ketoacidosis in diabetes. Proper dental hygiene, avoiding smoking, and following healthcare provider's instructions are recommended for treating bad breath. If bad breath persists or is accompanied by other symptoms, medical attention may be necessary.	Bad breath could be caused by poor oral hygiene, dental and gastroesophageal diseases, infections, smoking, alcoholism, and certain foods.
MEDIQA-MAS	Answer Summarization	For the following question, some relevant answers are given below. Please write down a short concise answer by summarizing the given answers. Question: Can I use a Power Plate machine if I have a pacemaker? Answer 1: Most machines and devices will not interfere with the pacemaker. But some with strong magnetic fields may. Always ask your provider about any specific device that you need to avoid. DO NOT put a magnet near the pacemaker. Most appliances in home are safe to be around. This includes ur refrigerator, washer, dryer, toaster, blender, computers and fax machines, hair dryer, stove, CD player, remote controls, and microwaves. You should keep several devices at least 12 inches (30 centimeters) away from the site where the pacemaker is placed under ur skin. These include: Battery powered cordless tools (such as screwdrivers and drills) Plug-in power tools (such as drills and table saws) Electric lawnmowers and leaf blowers Slot machines Stereo speakers Tell all providers that you have a pacemaker before any tests are done. Some medical equipment may interfere with ur pacemaker. Stay away from large motors, generators, and equipment. DO NOT lean over the open hood of a car that is running. Also stay away from: Radio transmitters and high-voltage power lines Products that use magnetic therapy, such as some mattresses, pillows, and massagers Large electrical- or gasoline-powered appliances If you have a cell phone: DO NOT put it in a pocket on the same side of the body as the pacemaker. When using cell phone, hold it to the ear on the opposite side of the body. Be careful around metal detectors and security wands. Handheld security wands may interfere with the pacemaker. Show the wallet card and ask to be hand searched. Most security gates at airports and stores are OK. But DO NOT stand near these devices for long periods. The pacemaker may set off alarms. After any operation, have the provider check the pacemaker. Answer 2: You will probably be able to go home after 1 day or even the same day in some cases. You should be able to return to your normal activity level quickly. Ask your provider how much you can use the arm on the side of the body where the pacemaker was placed. You may be advised not to: Lift anything heavier than 10 to 15 pounds (4.5 to 6.75 kilograms) Push, pull, and twist arm for 2 to 3 weeks. Raise arm above the shoulder for several weeks. When you leave the hospital, you will be given a card to keep in the wallet. This card lists the details of your pacemaker and has contact information for emergencies. You should always carry this wallet card with you. You should try to remember the name of the pacemaker manufacturer if you can in case you lose the card.	It is important to ask the provider about specific devices that may interfere with a pacemaker, but most appliances in the home are safe. It is recommended to stay away from large motors, generators, and equipment, products that use magnetic therapy, and to be cautious around metal detectors and security wands. Additionally, after getting a pacemaker, some activity restrictions may apply, such as not lifting heavy objects and limiting arm movements. It is important to carry a wallet card with pacemaker details and emergency contact information.	Most machines are safe to use with a pacemaker. Some devices should be kept at least 12 inches away. These include battery powered cordless tools (such as screwdrivers and drills) Plug-in power tools (such as drills and table saws) Electric lawnmowers and leaf blowers Slot machines Stereo speakers Some medical equipment may interfere with the pacemaker. Stay away from large motors, generators, and equipment, Radio transmitters and high-voltage power lines Products that use magnetic therapy, such as some mattresses, pillows, and massagers Large electrical- or gasoline-powered appliances. Do not put the cell phone in a pocket or to the ear on the same side as the pacemaker.

Table 23
Some examples of LLaMA-2-13b generated responses in Biomedical Literature Summarization (lay summary and abstract).

Dataset	Type	Prompt	LLaMA-2-13b Response	Gold Summary/Abstract
PLOS	Lay Summary	Write down a readable summary of the following biomedical article using less technical terminology (e.g., lay summary) such that it can be understandable for non-expert audiences: Naturally occurring genetic differences in gene regulation within populations underlie phenotypes of evolutionary and biomedical interest [1–3] and can serve as the basis for inference of regulatory networks [4,5] . A key problem in the field is understanding the molecular mechanisms by which DNA sequence variants give rise to expression change . Recent work has emphasized the importance of sequence differences in regions upstream of gene loci that harbor cis-acting determinants of transcription factor binding [6–10] and chromatin architecture [11–13]. Much less is known about the role of 3‘-end regulation as a determinant of expression variation between individuals . Alternative polyadenylation represents a major regulatory strategy in the human genome, with analysis across tissue types detecting multiple 3‘ UTR forms of over half of all human genes [14] . Detailed genetic studies have implicated polymorphisms affecting transcript termination in both Mendelian and complex human disease [15–20] . Genomic analyses have hinted at a broader role for genetic differences in RNA 3‘-end processing as a driver of expression variation [6,8,21,22], but the prevalence and the mechanisms of these changes are incompletely understood . Progress in dissecting the genetics of 3‘-end processing has been limited in part by fundamental questions about the regulatory information encoded in 3‘ UTRs . Single-gene studies have made clear that, in addition to its interplay with exonic splicing [23,24], RNA 3‘-end processing can dictate the extent of 3‘ UTR sequence incorporated into mature transcripts that governs half-life, translation, and localization [25–29]; the efficiency of transcription termination itself can also influence steady-state expression level of a given length form [30–34]. In general, however, identifying the regulatory elements that underlie relationships between 3‘ UTR sequence and gene expression remains a primary challenge, and for the majority of human genes, the regulatory impact of alternative polyadenylation is unknown . Likewise, the search for molecular players underlying cis-regulation of 3‘-end processing at individual gene loci [23] and genome-scale regulation of 3‘-end processing in trans [14,29,35–40] is an area of active research . A complete understanding of the genetics of alternative polyadenylation will require maps of transcript end site usage and 3‘ cis-regulatory elements, and analysis strategies to integrate the data . Recently developed short-read sequencing methods for transcript ends [40–46] have enabled the possibility of quantitative studies of the regulatory architecture of transcript end forms on a genomic scale . In this work, we set out to investigate mechanisms by which alternative polyadenylation impacts gene expression and its variation across genetically distinct human individuals . We used 3‘-end RNA-seq [42] to maximize the genomic coverage and precision of transcript end positions, and to measure quantitative expression levels of transcript forms . . . . These included variants attenuating usage of 3‘ forms of the translation initiation factor EIF2A and the putative DNA methylation enzyme DIP2B, as well as the expected effect of the polymorphic polyadenylation signal on usage of 3‘ forms of the inflammation.	The article discusses the importance of understanding the genetic basis of differences in gene expression between individuals. The authors present a study that used 3’ end RNA sequencing to investigate the role of genetic variation in the regulation of gene expression. They found that genetic differences in the 3’ untranslated regions (UTRs) of genes are associated with differences in the usage of alternative polyadenylation length forms, which in turn affects the stability and translation of mRNAs. The study identified specific motifs in the 3’ UTRs that are enriched in certain human populations and are associated with changes in gene expression. The findings suggest that genetic variation in the 3’ UTRs plays a significant role in the regulation of gene expression and may contribute to the development of complex diseases. In simpler terms, the study found that small changes in the DNA sequence of genes can affect how much of a specific protein is made in the body. These changes can affect the stability and translation of the mRNA, which carries the genetic information from the DNA to the ribosome where proteins are made. The study identified specific sequences in the mRNA that are more common in certain populations and are associated with changes in protein production. These findings suggest that small genetic changes in the mRNA can have a significant impact on the regulation of protein production and may contribute to the development of complex diseases.”	Messenger RNAs carry the instructions necessary to synthesize proteins that do work for the cell . Extending beyond the protein-coding sequence of a given mRNA is an additional stretch of sequence, harboring signals that govern how much protein is made and how long the mRNA remains in the cell before it is broken down . The incorporation of this end region into mature mRNA is itself subject to change; for the vast majority of human genes, how and why cells use different mRNA ends remains largely unknown . In this work, we surveyed mRNA ends from ~10, 000 genes in immune cells from genetically distinct human individuals . We found that mRNA end positions were not randomly distributed, but rather preferentially flanked the locations of regulatory signals that govern mRNA fate . The usage of these mRNA length forms and regulatory elements varied across individuals and could be dissected molecularly . Our results uncover key mechanisms and regulatory effects of transcript end processing, particularly as these are perturbed by genetic differences between humans .

(*continued on next page*)

Declaration of competing interest

The authors declare having no competing interests.

Declaration of Generative AI and AI-assisted technologies in the writing process

AI-assisted technologies are only used sometimes in the writing process to improve readability.

Acknowledgments

We would like to thank the handling editor and all the five reviewers for their excellent review comments. This research is supported by the research grant (RGPIN-2020-07157) from the Natural Sciences and Engineering Research Council (NSERC) of Canada, the York Research Chairs (YRC) program, and the generic research fund of York University, Canada. We also acknowledge Compute Canada for providing us with the computing resources to conduct experiments, as well as Anthropic for providing us early access to the Claude-2 API.

Table 23 (*continued*).

Dataset	Type	Prompt	LLaMA-2-13b Response	Gold Summary/Abstract
PLOS	Abstract	Write down the abstract of the following biomedical article: Naturally occurring genetic differences in gene regulation within populations underlie phenotypes of evolutionary and biomedical interest [1–3] and can serve as the basis for inference of regulatory networks [4,5]. A key problem in the field is understanding the molecular mechanisms by which DNA sequence variants give rise to expression change . Recent work has emphasized the importance of sequence differences in regions upstream of gene loci that harbor cis-acting determinants of transcription factor binding [6–10] and chromatin architecture [11–13]. Much less is known about the role of 3‘-end regulation as a determinant of expression variation between individuals . Alternative polyadenylation represents a major regulatory strategy in the human genome, with analysis across tissue types detecting multiple 3‘ UTR forms of over half of all human genes [14]. Detailed genetic studies have implicated polymorphisms affecting transcript termination in both Mendelian and complex human disease [15–20]. Genomic analyses have hinted at a broader role for genetic differences in RNA 3‘-end processing as a driver of expression variation [6,8,21,22], but the prevalence and the mechanisms of these changes are incompletely understood . Progress in dissecting the genetics of 3‘-end processing has been limited in part by fundamental questions about the regulatory information encoded in 3‘ UTRs . Single-gene studies have made clear that, in addition to its interplay with exonic splicing [23,24], RNA 3‘-end processing can dictate the extent of 3‘ UTR sequence incorporated into mature transcripts that governs half-life, translation, and localization [25–29]; the efficiency of transcription termination itself can also influence steady-state expression level of a given length form [30–34]. In general, however, identifying the regulatory elements that underlie relationships between 3‘ UTR sequence and gene expression remains a primary challenge, and for the majority of human genes, the regulatory impact of alternative polyadenylation is unknown . Likewise, the search for molecular players underlying cis-regulation of 3‘-end processing at individual gene loci [23] and genome-scale regulation of 3‘-end processing in trans [14,29,35–40] is an area of active research . A complete understanding of the genetics of alternative polyadenylation will require maps of transcript end site usage and 3‘ cis-regulatory elements, and analysis strategies to integrate the data . Recently developed short-read sequencing methods for transcript ends [40–46] have enabled the possibility of quantitative studies of the regulatory architecture of transcript end forms on a genomic scale . In this work, we set out to investigate mechanisms by which alternative polyadenylation impacts gene expression and its variation across genetically distinct human individuals . We used 3‘-end RNA-seq [42] to maximize the genomic coverage and precision of transcript end positions, and to measure quantitative expression levels of transcript forms . . . . These included variants attenuating usage of 3‘ forms of the translation initiation factor EIF2A and the putative DNA methylation enzyme DIP2B, as well as the expected effect of the polymorphic polyadenylation signal on usage of 3‘ forms of the inflammation.	The abstract of the article is: Naturally occurring genetic differences in gene regulation within populations underlie phenotypes of evolutionary and biomedical interest. Recent work has emphasized the importance of sequence differences in regions upstream of gene loci that harbor cis-acting determinants of transcription factor binding and chromatin architecture. However, the prevalence and mechanisms of these changes are incompletely understood. This study investigated the genetics of alternative polyadenylation, which can serve as a determinant of expression variation between individuals. The results shed light on the architecture of transcript ends and regulatory elements in human 3’ UTRs and the principles of genetic variation in 3’ length form usage. The study found that alternative polyadenylation reaches a threshold of detection in approximately 30% of genes, with the majority of reads originating from 3’ UTRs of coding genes. The breakdown of transcript forms into classes revealed that alternative polyadenylation is enriched in genes with immune-related functions, highlighting the importance of 3’ end processing in the control of gene expression levels for specialized cell functions. The study also found that genetic differences in RNA 3’ end processing underlie changes across individuals in steady-state levels of gene expression, and that naturally occurring genetic variation in polyadenylation signals can underlie changes in expression levels. The study used 3’ end RNA-seq to maximize the genomic coverage and precision of transcript end positions and to measure quantitative expression levels of transcript forms. The results provide a complete understanding of the genetics of alternative polyadenylation and its impact on gene expression and its variation across genetically distinct human individuals.	Gene expression varies widely between individuals of a population, and regulatory change can underlie phenotypes of evolutionary and biomedical relevance . A key question in the field is how DNA sequence variants impact gene expression, with most mechanistic studies to date focused on the effects of genetic change on regulatory regions upstream of protein-coding sequence . By contrast, the role of RNA 3‘-end processing in regulatory variation remains largely unknown, owing in part to the challenge of identifying functional elements in 3‘ untranslated regions . In this work, we conducted a genomic survey of transcript ends in lymphoblastoid cells from genetically distinct human individuals . Our analysis mapped the cis-regulatory architecture of 3‘ gene ends, finding that transcript end positions did not fall randomly in untranslated regions, but rather preferentially flanked the locations of 3‘ regulatory elements, including miRNA sites . The usage of these transcript length forms and motifs varied across human individuals, and polymorphisms in polyadenylation signals and other 3‘ motifs were significant predictors of expression levels of the genes in which they lay . Independent single-gene experiments confirmed the effects of polyadenylation variants on steady-state expression of their respective genes, and validated the regulatory function of 3‘ cis-regulatory sequence elements that mediated expression of these distinct RNA length forms . Focusing on the immune regulator IRF5, we established the effect of natural variation in RNA 3‘-end processing on regulatory response to antigen stimulation . The results underscore the importance of two mechanisms at play in the genetics of 3‘-end variation: the usage of distinct 3‘-end processing signals and the effects of 3‘ sequence elements that determine transcript fate . The findings suggest that the strategy of integrating observed 3‘-end positions with inferred 3‘ regulatory motifs will prove to be a critical tool in continued efforts to interpret human genome variation.

Appendix

A.1. Performance based on other metrics

To provide further insights on the performance of different zero-shot LLMs, the results based on the accuracy metric are demonstrated in Table 19. From Table 19, it can be observed that while zero-shot LLMs obtain poor performance in the NER datasets in terms of precision, recall, and F1 (see Table 9), they obtain very high performance in terms of the accuracy metric. This is because in the biomedical dataset, there are a large number of entities that are not related to any biomedical tags (i.e., ‘O’ tags). Thus, LLMs may obtain higher accuracy because of predicting non-biomedical entities more accurately rather than identifying the required biomedical entities. Meanwhile, in the text classification datasets, the performance of LLMs in terms of accuracy is quite bad similar to their original results in terms of the F1 metric (see Table 8).

A.2. Average performance per task

To provide further insights, we demonstrate the average performance of different LLMs across each task in Fig. 4.

A.3. Sample response generated by LLMs

Some examples of the responses generated by different LLMs in various tasks are demonstrated as follows:

— See Table 20 for some example responses of PaLM-2 in Relation Extraction and Text Classification tasks.

— See Table 21 for some example responses of Claude-2 in Entity Linking, Named Entity Recognition, and Question Answering tasks.

— See Table 22 for some example responses of GPT-3.5 in various Summarization tasks: Dialogue Summarization, Question Summarization, and Answer Summarization.

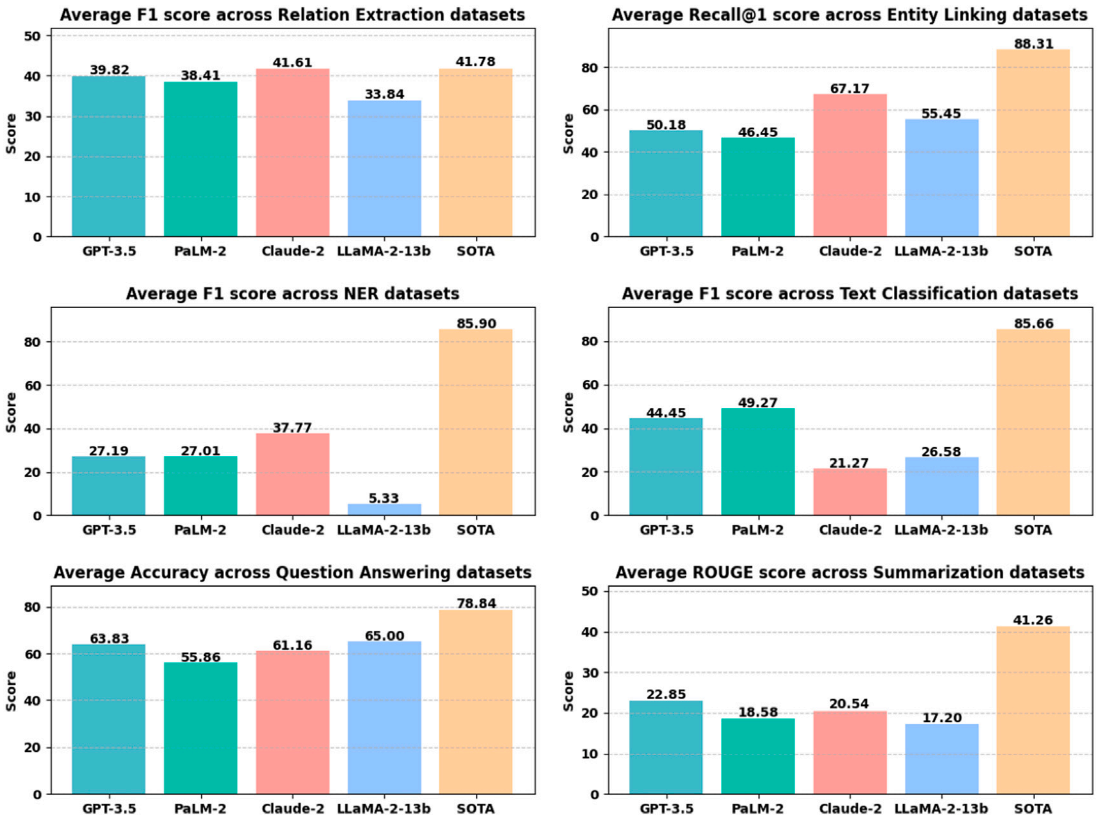


Fig. 4. Average score by different zero-shot LLMs across each task.

— See Table 23 for some example responses of LLaMA-2-13b in Biomedical Literature Summarization.

References

[1] Anna Rogers, Olga Kovaleva, Anna Rumshisky, A primer in BERTology: What we know about how BERT works, Trans. Assoc. Comput. Linguist. 8 (2021) 842–866.

[2] Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, Sivanesan Sangeetha, AMMU: a survey of transformer-based biomedical pretrained language models, J. Biomed. Inform. 126 (2022) 103982.

[3] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (4) (2020) 1234–1240.

[4] Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, Sheng Yu, BioBART: Pretraining and evaluation of a biomedical generative language model, in: Proceedings of the 21st Workshop on Biomedical Language Processing, 2022, pp. 97–109.

[5] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, Tie-Yan Liu, BioGPT: generative pre-trained transformer for biomedical text generation and mining, Brief. Bioinform. 23 (6) (2022).

[6] Wei Wang, Vincent W. Zheng, Han Yu, Chunyan Miao, A survey of zero-shot learning: Settings, methods, and applications, ACM Trans. Intell. Syst. Technol. 10 (2) (2019) 1–37.

[7] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al., A survey of large language models, 2023, arXiv preprint arXiv:2303.18223.

[8] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, Graham Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, ACM Comput. Surv. 55 (9) (2023) 1–35.

[9] Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, Jimmy Huang, Evaluation of chatGPT on biomedical tasks: A zero-shot comparison with fine-tuned generative transformers, in: The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 326–336.

[10] Zhaoshan Liu, Qiujie Lv, Ziduo Yang, Yifan Li, Chau Hung Lee, Lei Shen, Recent progress in transformer-based medical image analysis, Comput. Biol. Med. (2023) 107268.

[11] Tawsifur Rahman, Amith Khandakar, Yazan Qiblawey, Anas Tahir, Serkan Kiranyaz, Saad Bin Abul Kashem, Mohammad Tariqul Islam, Somaya Al Maadeed, Susu M Zughaier, Muhammad Salman Khan, et al., Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images, Comput. Biol. Med. 132 (2021) 104319.

[12] Mohammad Amin Morid, Alireza Borjali, Guilherme Del Fiol, A scoping review of transfer learning research on medical image analysis using ImageNet, Comput. Biol. Med. 128 (2021) 104115.

[13] Aaron M. Cohen, William R. Hersh, A survey of current work in biomedical text mining, Brief. Bioinform. 6 (1) (2005) 57–71.

[14] Benyou Wang, Qianqian Xie, Jiahuan Pei, Zhihong Chen, Prayag Tiwari, Zhao Li, Jie Fu, Pre-trained language models in biomedical domain: A systematic survey, ACM Comput. Surv. (2021).

[15] Jacob O'Brien, Heyam Hayder, Yara Zayed, Chun Peng, Overview of microRNA biogenesis, mechanisms of actions, and circulation, Front. Endocrinol. 9 (2018) 402.

[16] Yanrong Ji, Zhihan Zhou, Han Liu, Ramana V. Davuluri, DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome, Bioinformatics 37 (15) (2021) 2112–2120.

[17] Md Mamun Ali, Bikash Kumar Paul, Kawsar Ahmed, Francis M Bui, Julian MW Quinn, Mohammad Ali Moni, Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison, Comput. Biol. Med. 136 (2021) 104672.

[18] Bilal Shaker, Sajjad Ahmad, Jingyu Lee, Chanjin Jung, Dokyun Na, In silico methods and tools for drug discovery, Comput. Biol. Med. 137 (2021) 104851.

[19] Dominic D. Martinelli, Generative machine learning for de novo drug discovery: A systematic review, Comput. Biol. Med. 145 (2022) 105403.

[20] Sanjeevi Pandiyan, Li Wang, A comprehensive review on recent approaches for cancer drug discovery associated with artificial intelligence, Comput. Biol. Med. (2022) 106140.

[21] Vu Hong Loan Nguyen, Rebecca Hough, Stefanie Bernaudo, Chun Peng, Wnt/β-catenin signalling in ovarian cancer: Insights into its hyperactivation and function in tumorigenesis, J. Ovarian Res. 12 (2019) 1–17.

[22] Safoura Soleymani, Amin Tavassoli, Mohammad Reza Housaindokht, An overview of progress from empirical to rational design in modern vaccine development, with an emphasis on computational tools and immunoinformatics approaches, Comput. Biol. Med. 140 (2022) 105057.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, Illia Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 4171–4186.

[25] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, Language models are unsupervised multitask learners, OpenAI Blog 1 (8) (2019) 9.

[26] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, Matthew McDermott, Publicly available clinical BERT embeddings, 2019, arXiv preprint arXiv:1904.03323.

[27] Iz Beltagy, Kyle Lo, Arman Cohan, SciBERT: A pretrained language model for scientific text, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP, 2019, pp. 3615–3620.

[28] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, Hoifung Poon, Domain-specific language model pretraining for biomedical natural language processing, 2020, arXiv preprint arXiv:2007.15779.

[29] Yifan Peng, Shankai Yan, Zhiyong Lu, Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets, in: Proceedings of the 18th BioNLP Workshop and Shared Task, 2019, pp. 58–65.

[30] Kamal raj Kanakarajan, Bhuvana Kundumani, Malaikannan Sankarasubbu, BioELECTRA: pretrained biomedical text encoder using discriminators, in: Proceedings of the 20th Workshop on Biomedical Language Processing, 2021, pp. 143–154.

[31] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019, arXiv preprint arXiv:1910.13461.

[32] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, Hoifung Poon, Domain-specific language model pretraining for biomedical natural language processing, ACM Trans. Comput. Healthc. 3 (1) (2021) 1–23.

[33] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al., Language models are few-shot learners, 2020, arXiv preprint arXiv:2005.14165.

[34] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al., Training language models to follow instructions with human feedback, Adv. Neural Inf. Process. Syst. 35 (2022) 27730–27744.

[35] Leslie Pack Kaelbling, Michael L. Littman, Andrew W. Moore, Reinforcement learning: A survey, J. Artif. Intell. Res. 4 (1996) 237–285.

[36] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al., Palm: Scaling language modeling with pathways, 2022, arXiv preprint arXiv:2204.02311.

[37] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al., Palm 2 technical report, 2023, arXiv preprint arXiv:2305.10403.

[38] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al., Large language models encode clinical knowledge, Nature (2023) 1–9.

[39] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al., Llama: Open and efficient foundation language models, 2023, arXiv preprint arXiv:2302.13971.

[40] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, 2023, arXiv preprint arXiv:2307.09288.

[41] Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, Diyi Yang, Is ChatGPT a general-purpose natural language processing task solver? 2023, arXiv preprint arXiv:2302.06476.

[42] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al., A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity, 2023, arXiv preprint arXiv:2302.04023.

[43] Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, Wei Cheng, Exploring the limits of ChatGPT for query or aspect-based text summarization, 2023, arXiv preprint arXiv:2302.08081.

[44] Vikas Yadav, Steven Bethard, A survey on recent advances in named entity recognition from deep learning models, in: Emily M. Bender, Leon Derczynski, Pierre Isabelle (Eds.), Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 2145–2158, URL https://aclanthology.org/C18-1182.

[45] Zexuan Zhong, Danqi Chen, A frustratingly easy approach for entity and relation extraction, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 50–61.

[46] Md Tahmid Rahman Laskar, Cheng Chen, Jonathan Johnston, Xue-Yong Fu, Shashi Bhushan TN, Simon Corston-Oliver, An auto encoder-based dimensionality reduction technique for efficient entity linking in business phone conversations, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 3363–3367.

[47] Md Tahmid Rahman Laskar, Cheng Chen, Aliaksandr Martsinovich, Jonathan Johnston, Xue-Yong Fu, Shashi Bhushan Tn, Simon Corston-Oliver, BLINK with elasticsearch for efficient entity linking in business conversations, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track, 2022, pp. 344–352.

[48] Asma Ben Abacha, Chaitanya Shivade, Dina Demner-Fushman, Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering, in: Proceedings of the 18th BioNLP Workshop and Shared Task, 2019, pp. 370–379.

[49] Yun He, Ziwei Zhu, Yin Zhang, Qin Chen, James Caverlee, Infusing disease knowledge into BERT for health question answering, medical inference and disease name recognition, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, Association for Computational Linguistics, 2020, pp. 4604–4614, Online.

[50] Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, Hoda K Mohamed, Automatic text summarization: A comprehensive survey, Expert Syst. Appl. 165 (2021) 113679.

[51] Larry Smith, Lorraine K Tanabe, Cheng-Ju Kuo, I Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, Kuzman Ganchev, Manabu Torii, et al., Overview of BioCreative II gene mention recognition, Genome Biol. 9 (2) (2008) 1–19.

[52] Nigel Collier, Jin-Dong Kim, Introduction to the bio-entity recognition task at JNLPBA, in: Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications, NLPBA/BioNLP, 2004, pp. 73–78.

[53] Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, et al., The CHEMDNER corpus of chemicals and drugs and its annotation principles, J. Cheminformatics 7 (1) (2015) 1–17.

[54] Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, Zhiyong Lu, BioCreative V CDR task corpus: a resource for chemical disease relation extraction, Database 2016 (2016).

[55] Rezarta Islamaj Doğan, Robert Leaman, Zhiyong Lu, NCBI disease corpus: a resource for disease name recognition and concept normalization, J. Biomed. Inform. 47 (2014) 1–10.

[56] Martin Gerner, Goran Nenadic, Casey M. Bergman, LINNAEUS: a species name identification system for biomedical literature, BMC Bioinform. 11 (1) (2010) 1–17.

[57] Evangelos Pafilis, Sune P Frankild, Lucia Fanini, Sarah Faulwetter, Christina Pavloudi, Aikaterini Vasileiadou, Christos Arvanitidis, Lars Juhl Jensen, The SPECIES and ORGANISMS resources for fast and accurate identification of taxonomic names in text, PLoS One 8 (6) (2013) e65390.

[58] Yutai Hou, Yingce Xia, Lijun Wu, Shufang Xie, Yang Fan, Jinhua Zhu, Tao Qin, Tie-Yan Liu, Discovering drug–target interaction knowledge from biomedical literature, Bioinformatics 38 (22) (2022) 5100–5107.

[59] María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, Thierry Declerck, The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions, J. Biomed. Inform. 46 (5) (2013) 914–920.

[60] Hongyi Yuan, Zheng Yuan, Sheng Yu, Generative biomedical entity linking via knowledge base-guided pre-training and synonyms-aware fine-tuning, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022, pp. 4038–4048.

[61] Marco Basaldella, Fangyu Liu, Ehsan Shareghi, Nigel Collier, COMETA: A corpus for medical entity linking in the social media, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, Association for Computational Linguistics, 2020, pp. 3122–3137, Online.

[62] Simon Baker, Ilona Silins, Yufan Guo, Imran Ali, Johan Högberg, Ulla Stenius, Anna Korhonen, Automatic semantic classification of scientific literature according to the hallmarks of cancer, Bioinformatics 32 (3) (2016) 432–440.

[63] Qingyu Chen, Alexis Allot, Zhiyong Lu, LitCovid: an open database of COVID-19 literature, Nucleic Acids Res. 49 (D1) (2021) D1534–D1540.

[64] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, Xinghua Lu, Pubmedqa: A dataset for biomedical research question answering, 2019, arXiv preprint arXiv:1909.06146.

[65] Md Tahmid Rahman Laskar, Xiangji Huang, Enamul Hoque, Contextualized embeddings based transformer encoder for sentence similarity modeling in answer selection task, in: Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pp. 5505–5514.

[66] Asma Ben Abacha, Dina Demner-Fushman, On the summarization of consumer health questions, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 2228–2234.

[67] Asma Ben Abacha, Yassine M'rabet, Yuhao Zhang, Chaitanya Shivade, Curtis Langlotz, Dina Demner-Fushman, Overview of the MEDIQA 2021 shared task on summarization in the medical domain, in: Proceedings of the 20th Workshop on Biomedical Language Processing, 2021, pp. 74–85.

[68] Max Savery, Asma Ben Abacha, Soumya Gayen, Dina Demner-Fushman, Question-driven summarization of answers to consumer health questions, Sci. Data 7 (1) (2020) 1–9.

[69] Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, et al., MedDialog: Large-scale medical dialogue datasets, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2020, pp. 9241–9250.

[70] Khalil Mrini, Franck Dernoncourt, Seunghyun Yoon, Trung Bui, Walter Chang, Emilia Farcas, Ndapandula Nakashole, A gradually soft multi-task and data-augmented approach to medical question understanding, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 1505–1515.

[71] Zheheng Luo, Qianqian Xie, Sophia Ananiadou, Readability controllable biomedical document summarization, in: Findings of the Association for Computational Linguistics, EMNLP 2022, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 4667–4680.

[72] Tomas Goldsack, Zhihao Zhang, Chenghua Lin, Carolina Scarton, Making science simple: Corpora for the lay summarisation of scientific literature, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 10589–10604.

[73] Tomas Goldsack, Zheheng Luo, Qianqian Xie, Carolina Scarton, Matthew Shardlow, Sophia Ananiadou, Chenghua Lin, BioLaySumm 2023 shared task: Lay summarisation of biomedical research articles, in: The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, 2023, pp. 468–477.

[74] OpenAI, GPT-4 technical report, 2023.

[75] Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, Jimmy Huang, A systematic study and comprehensive evaluation of chatGPT on benchmark datasets, in: Findings of the Association for Computational Linguistics, ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 431–469.

[76] Chin-Yew Lin, Rouge: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, 2004, pp. 74–81.

[77] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, Yoav Artzi, BERTScore: Evaluating text generation with BERT, in: International Conference on Learning Representations, 2019.

[78] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, Radu Soricut, ALBERT: A lite BERT for self-supervised learning of language representations, in: International Conference on Learning Representations, 2019.

[79] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al., Scaling instruction-finetuned language models, 2022, arXiv preprint arXiv:2210.11416.

[80] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al., Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (140) (2020) 1–67.

[81] Wen Xiao, Iz Beltagy, Giuseppe Carenini, Arman Cohan, PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization, in: Smaranda Muresan, Preslav Nakov, Aline Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 5245–5263, http://dx.doi.org/10.18653/v1/2022.acl-long.360, URL https://aclanthology.org/2022.acl-long.360.

[82] Iz Beltagy, Matthew E. Peters, Arman Cohan, Longformer: The long-document transformer, 2020, arXiv preprint arXiv:2004.05150.

[83] Bernal Jiménez Gutiérrez, Jucheng Zeng, Dongdong Zhang, Ping Zhang, Yu Su, Document classification for COVID-19 literature, in: Findings of the Association for Computational Linguistics, EMNLP 2020, 2020, pp. 3715–3722.

[84] Mong Yuan Sim, Xiang Dai, Maciej Rybinski, Sarvnaz Karimi, CSIRO data61 team at BioLaySumm task 1: Lay summarisation of biomedical research articles using generative models, in: The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, 2023, pp. 629–635.

[85] Chao-Yi Chen, Jen-Hao Yang, Lung-Hao Lee, NCUEE-NLP at BioLaySumm task 2: Readability-controlled summarization of biomedical articles using the PRIMERA models, in: The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, 2023, pp. 586–591.

[86] Md Tahmid Rahman Laskar, Enamul Hoque, Jimmy Xiangji Huang, Domain adaptation with pre-trained transformers for query-focused abstractive text summarization, Comput. Linguist. 48 (2) (2022) 279–320.

[87] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov, RoBERTa: A robustly optimized bert pretraining approach, 2019, arXiv preprint arXiv:1907.11692.

[88] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, Percy Liang, Lost in the middle: How language models use long contexts, 2023, arXiv preprint arXiv:2307.03172.

[89] Md Tahmid Rahman Laskar, Xue-Yong Fu, Cheng Chen, Shashi Bhushan Tn, Building real-world meeting summarization systems using large language models: A practical perspective, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track, 2023, pp. 343–352.

[90] Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, et al., A comprehensive capability analysis of gpt-3 and gpt-3.5 series models, 2023, arXiv preprint arXiv:2303.10420.

[91] Changmao Li, Jeffrey Flanigan, Task contamination: Language models may not be few-shot anymore, 2023, arXiv preprint arXiv:2312.16337.

[92] Shaoxiong Ji, Matti Hölttä, Pekka Marttinen, Does the magic of BERT apply to medical code assignment? A quantitative study, Comput. Biol. Med. 139 (2021) 104998.

[93] Nelson RC Monteiro, Tiago O Pereira, Ana Catarina D Machado, José L Oliveira, Maryam Abbasi, Joel P Arrais, FSM-DDTR: End-to-end feedback strategy for multi-objective de novo drug design using transformers, Comput. Biol. Med. 164 (2023) 107285.

[94] Syed Muazzam Ali Shah, Semmy Wellem Taju, Quang-Thai Ho, Yu-Yen Ou, et al., GT-finder: Classify the family of glucose transporters with pre-trained BERT language models, Comput. Biol. Med. 131 (2021) 104259.

[95] Long Phan, Tai Dang, Hieu Tran, Trieu Trinh, Vy Phan, Lam Chau, Minh-Thang Luong, Enriching biomedical knowledge for low-resource language through large-scale translation, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 2023, pp. 3123–3134.

[96] Xiangji Huang, Ming Zhong, Luo Si, York university at TREC 2005: Genomics track, in: Ellen M. Voorhees, Lori P. Buckland (Eds.), Proceedings of the Fourteenth Text REtrieval Conference, TREC 2005, Gaithersburg, Maryland, USA, November 15–18, 2005, in: NIST Special Publication, vol. 500–266, National Institute of Standards and Technology (NIST), 2005.

[97] Xiangji Huang, Qinmin Hu, A bayesian learning approach to promoting diversity in ranking for biomedical information retrieval, in: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2009, pp. 307–314.

[98] Xiaoshi Yin, Jimmy Xiangji Huang, Xiaofeng Zhou, Zhoujun Li, A survival modeling approach to biomedical search result diversification using wikipedia, in: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2010, pp. 901–902.

[99] Nazish Khalid, Adnan Qayyum, Muhammad Bilal, Ala Al-Fuqaha, Junaid Qadir, Privacy-preserving artificial intelligence in healthcare: Techniques and applications, Comput. Biol. Med. (2023) 106848.

[100] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al., Gemini: a family of highly capable multimodal models, 2023, arXiv preprint arXiv:2312.11805.

[101] Ruibo Chen, Tianyi Xiong, Yihan Wu, Guodong Liu, Zhengmian Hu, Lichang Chen, Yanshuo Chen, Chenxi Liu, Heng Huang, GPT-4 vision on medical image classification–A case study on COVID-19 dataset, 2023, arXiv preprint arXiv:2310.18498.

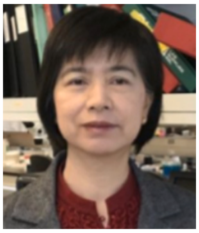
[102] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, et al., Large-scale domain-specific pretraining for biomedical vision-language processing, 2023, arXiv preprint arXiv:2303.00915.

[103] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, Weidi Xie, Pmc-vqa: Visual instruction tuning for medical visual question answering, 2023, arXiv preprint arXiv:2305.10415.

[104] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, Pranav Rajpurkar, Med-flamingo: a multimodal medical few-shot learner, in: Machine Learning for Health (ML4H), PMLR, 2023, pp. 353–367.

Israt Jahan is currently pursuing her Master's degree in Biology at York University. She has published several peer-reviewed papers in referred biomedical journals and conference proceedings, which include Frontiers in Plant Science, BioNLP, 2023, etc. She is currently working on utilizing Large Language Models for biomedical tasks. Israt also has prior research experience in Molecular Biology, with a focus on Genetics and Plant Breeding.

Md Tahmid Rahman Laskar is an Applied Scientist at Dialpad Canada Inc., and a Course Director (Adjunct Faculty) at York University. Prior to that, he obtained his Master's degree in Computer Science from York University. He has published research papers in several reputed journals, such as Computational Linguistics, ACM Transactions on Cyber–Physical Systems, Computers in Human Behavior, ACM Computing Surveys, and the Journal of Ambient Intelligence and Humanized Computing. In addition, he has published research papers in reputed conference proceedings, such as ACL, SIGIR, EMNLP, NAACL, COLING, etc. He was also awarded the best paper award at DaSH@EMNLP 2022. Tahmid is currently interested on optimizing Large Language Models in terms of cost, efficiency, and factual correctness.



Dr. Chun Peng is a professor of Biology at York University and a Tier 1 York Research Chair in Women's Reproductive Health. She received her PhD from the University of Alberta in 1993 and joined York University as an assistant professor in 1995. Her group studies how signaling molecules control ovarian and placenta development and how their dysregulation contributes to the pathogenesis of ovarian cancer and pregnancy complications, such as preeclampsia. Dr. Peng has published over 130 articles in international journals, including Cancer Research, Oncogene, Molecular Therapy, and Nature Communication, with over 12,000 citations. Her research is funded by Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, and Cancer Research Society.

Professor Jimmy Huang is holding a York Research Chair Professorship in big data analytics. His major research focuses on the areas of information retrieval, artificial intelligence, and big data analytics with complex structures and their applications to the Web and medical healthcare. He has published over 300 refereed papers in top-tier journals (such as ACM Transactions on Information Systems, Journal of American Society for Information Science & Technology, IEEE Transactions on Knowledge & Data Engineering, IEEE/ACM Transactions on Audio, Speech and Language Processing, ACM Transactions on Intelligent Systems and Technology, Information Sciences, Information Retrieval, BMC Bioinformatics, BMC Medical Genomics and BMC Genomics) and premier conferences in the fields (such as ACM SIGIR, ACM CIKM, ACM SIGKDD, COLING, ACL, IJCAI and AAAI). He was the General Conference Chair for the 19th International ACM CIKM Conference and the General Conference Chair for the 43rd International ACM SIGIR Conference.