STUDYING THE EFFECTIVENESS OF LARGE LANGUAGE MODELS IN BENCHMARK BIOMEDICAL TASKS

ISRAT JAHAN

A DISSERTATION SUBMITTED TO THE FACULTY OF GRADUATE STUDIES IN

PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF MASTER OF SCIENCE

GRADUATE PROGRAM IN BIOLOGY

YORK UNIVERSITY

TORONTO, ONTARIO, CANADA

JULY, 2024

©ISRAT JAHAN, 2024

Abstract

Recently, Large Language Models (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 thesis 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. Interestingly, this evaluation shows 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 pretraining on large text corpora makes LLMs quite specialized even in the biomedical domain. The findings also shows 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, this study demonstrates that LLMs have the potential to be a valuable tool for various biomedical tasks that lack large annotated data.

Dedication

This thesis is dedicated to my parents, whose boundless love, unwavering support, and immeasurable sacrifices have laid the foundation for all my achievements. Their guidance and belief in my potential have been the pillars that have sustained me throughout this journey.

I also dedicate this work to my beloved husband, whose constant support, patience, and encouragement have been invaluable. His unwavering faith and dedication have been crucial in guiding me through the challenges and triumphs of this work.

Above all, I express my deepest gratitude to Allah, whose infinite mercy and guidance have sustained me throughout this endeavor. Without His grace, none of this would have been possible.

Acknowledgement

In the name of the Almighty Allah, the Most Merciful, all glory and praise be to Him. It is solely due to His profound blessings that I have been able to successfully complete my thesis-based Master's degree in Biology at York University.

I wish to express my deepest gratitude to a number of people and organizations for their unwavering academic, financial, and emotional support throughout my Master's program. First, I would like to thank my supervisors, **Dr. Jimmy Xiangji Huang** and **Dr. Chun Peng** for giving me the opportunity to work under their kind supervision. I express my profound gratitude to Dr. Jimmy Xiangji Huang for his steadfast guidance, meticulous advice, and unwavering encouragement throughout my research endeavor. His scholarly input, constant support, and affectionate approach were instrumental in the successful conclusion of this research. I wish to extend my heartfelt appreciation to Dr. Chun Peng for her indispensable advice, innovative suggestions, constructive critiques, and essential guidance throughout the research.

I would also like to thank **Dr. Sapna Sharma** for being a part of my supervisory committee. I appreciate her continuous interest and involvement in my research as well as spending her valuable time on my thesis.

I extend my sincere gratitude to Dr. Jimmy Xiangji Huang's former M.Sc. and forthcoming Ph.D. student, **Md Tahmid Rahman Laskar**, for his invaluable mentorship and assistance in acquainting me with the intricacies of Large Language Models for my research

endeavors. His guidance has been instrumental in navigating the complexities of my work and advancing my understanding of this field. I deeply appreciate his support and guidance, which have been indispensable to my research progress.

This research is supported by the Natural Sciences & Engineering Research Council (NSERC) of Canada. The experiments with Claude-2 are supported by the free API credits given to some York University researchers by Anthropic. Moreover, the computing resources to run experiments using the LLaMA-2-13B model are supported by Compute Canada. A big thanks to them as well.

Finally, I wish to extend my deepest appreciation to my parents, elder sister, dear husband, and in-laws for their unwavering love and support, which have been essential in propelling me forward. Their constant encouragement and belief in me have provided the strength, motivation, and optimism needed during challenging times. In particular, I am immensely grateful to my mother for her selfless sacrifices and heartfelt prayers which have been a source of immense comfort and inspiration.

Preface

This thesis is submitted to the Faculty of Graduate Studies in partial fulfillment of the requirements for a Master of Science Degree in Biology. The work presented here is done by the author Israt Jahan under the supervision of Dr. Jimmy Xiangji Huang and Dr. Chun Peng. Some parts of this thesis have been published in the Computers in Biology and Medicine journal[1] (Jahan et al., 2024), and the 22nd Workshop[2] on Biomedical Natural Language Processing and BioNLP Shared Tasks (Jahan et al., 2023).

[1]Impact Factor 7.7 (as of June 10, 2024). Peer-reviewed by 6 reviewers.
[2]Peer-reviewed by 3 reviewers.

List of Publications from the Thesis

- Jahan, I., Laskar, M. T. R., Peng, C., & Huang, J. X. (2024, March). A Comprehensive Evaluation of Large Language Models on Benchmark Biomedical Text Processing Tasks. Computers in Biology and Medicine, 108189. https://www.sciencedirect.com/science/article/pii/S0010482524002737

- Jahan, I., Laskar, M. T. R., Peng, C., & Huang, J. X. (2023, July). 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 (pp. 326-336). https://aclanthology.org/2023.bionlp-1.30/

Table of Contents

Page

Abstract **ii**

Dedication **iii**

Acknowledgement **v**

Preface **vi**

List of Publications **vii**

List of Tables **xiii**

List of Abbreviations **xviii**

1 Introduction **1**

1.1 Background 2

1.1.1 Language Models for The Biomedical Domain 3

1.1.2 Large Language Models 4

1.1.3 Biomedical Tasks Description 7

1.1.4 Prompting Large Language Models 11

1.2 Hypothesis 13
1.3 Research Objectives 14

2 Methodology **16**
2.1 Datasets and Prompts 17
2.1.1 *NER* 17
2.1.2 *Relation extraction* 19
2.1.3 *Entity linking* 20
2.1.4 *Question answering* 20
2.1.5 *Text classification* 21
2.1.6 *Text summarization* 21
2.2 Models 24
2.2.1 *GPT-3.5* 25
2.2.2 *PaLM-2* 25
2.2.3 *Claude-2* 26
2.2.4 *LLaMA-2* 26
2.3 Evaluation Pipeline 27
2.4 Evaluation Metrics 29
2.5 Baselines 31

3 Results and Discussion **34**
3.1 Performance Comparisons 34
3.1.1 *Relation extraction* 34
3.1.2 *Text classification* 36
3.1.3 *Question answering* 36
3.1.4 *Entity linking* 38

3.1.5 *NER* . . . 39
3.1.6 *Summarization* . . . 40
3.2 Analyses . . . 44
3.2.1 *Effects of prompt variation* . . . 44
3.2.2 *Effects of few-shot learning* . . . 44
3.2.3 Effects of Fine-Tuning . . . 47
3.2.4 *Data contamination detection analysis* . . . 49
3.2.5 *Average performance per task* . . . 52
3.2.6 Cost Analysis . . . 53
3.3 Key Findings . . . 54

4 Conclusions and Future Work **56**
4.1 Contributions of The Research . . . 56
4.2 Limitations of The Research . . . 57
4.3 Concluding Remarks and Future Directions . . . 58

Bibliography **60**

A Appendix **82**

List of Figures

1.1 An overview of various pre-training techniques depending on the model type: on the left, (i) BERT for the encoder-based models to predict the masked words/tokens in a sequence (e.g., masked language modeling for text classification) based on the bidirectional context, then (ii) GPT for the decoder-based model to predict the next word in a sequence (e.g., auto-regressive language modeling for text generation) based on the context from the previous words, and finally (iii) BART for the encoder-decoder based model that utilizes both the encoder and the decoder of transformer with the encoders being used for text classification type objective while the decoders being used for text generation objectives. This figure is taken from the BART paper (Lewis et al., 2019). 5

1.2 Overview of the continual pre-training of language models (e.g., BioBERT) for the adaptation to the biomedical domain followed by fine-tuning. This figure is also taken from the BioBERT paper (Lee et al., 2020). 6

1.3 An overview of the overall training procedure of LLMs that combines supervised learning with reinforcement learning to train a reward model. This figure is taken from OpenAI (https://openai.com/index/chatgpt/). Last accessed June 1, 2024. 7

1.4 Overview of zero-shot, one-shot, few-shot, and traditional fine-tuning techniques. This figure is taken from the GPT-3 paper (Brown et al., 2020). 11

2.1 An overview of the methodology to evaluate 6 biomedical tasks across 26 datasets in this thesis. At first, the prompts are constructed for each dataset. Then, the responses are generated for each dataset using respective LLMs. Finally, depending on the task, various evaluation techniques are applied. 16

3.1 Confusion Matrix for different models in the PubMedQA dataset. 37

3.2 Confusion Matrix for different models in the MediQA-2019 dataset. 38

3.3 Average score by different zero-shot LLMs across each task. 52

List of Tables

2.1 Sample Prompts in Different Named Entity Recognition (NER) Datasets. . . . 18

2.2 Sample Prompts in Different Relation Extraction Datasets. 19

2.3 Sample Prompts in Different Entity Linking Datasets. 20

2.4 Sample Prompts in Different Question Answering Datasets. 21

2.5 Sample Prompts in Different Text Classification Datasets. 22

2.6 Sample Prompts in Different Text Summarization tasks. 24

3.1 Performance on Relation Extraction datasets. All SOTA results are taken from the BioGPT (Luo et al., 2022a) model. 35

3.2 Performance on Text Classification, Question Answering, and Entity Linking datasets. The SOTA results for HoC and PubMedQA are taken from the BioGPT (Luo et al., 2022a) model, while for LitCovid and MediQA-2019 datasets, the SOTA results are taken from Gutiérrez et al. (2020) and He et al. (2020), respectively. Note that all SOTA results for Entity Linking are taken from the BioBART (Yuan et al., 2022a) model. 36

3.3 Performance on Named Entity Recognition datasets. SOTA results are from the BioBERT (Lee et al., 2020) model. Here, 'Precision' and 'Recall' are denoted by 'P' and 'R', respectively. 39

3.4 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. The SOTA (State-of-the-Art) results are taken from the BioBART (Yuan et al., 2022a) model. Also, LLaMA-2 refers to its 13b version, similar to other tasks. 40
3.5 Performance on the Biomedical Lay Summarization task. SOTA results are from Sim et al. (2023). 41
3.6 Performance on Readability Controlled Summarization in the PLOS dataset. SOTA results are from Chen et al. (2023a). 41
3.7 Performance of different LLMs on Biomedical Lay Summarization datasets based on various input lengths. Here, ‘R-1’, ‘R-2’, ‘R-L’ and ‘B-S’ denote ‘ROUGE-1’, ‘ROUGE-2’, ‘ROUGE-L’, and ‘BERTScore’, respectively. 43
3.8 Performance of different LLMs on Readability Controlled Summarization in the PLOS dataset based on various input lengths. Here, ‘R-1’, ‘R-2’, ‘R-L’ and ‘B-S’ denote ‘ROUGE-1’, ‘ROUGE-2’, ‘ROUGE-L’, and ‘BERTScore’, respectively. 43
3.9 Effects of Prompt Variations in GPT-3.5 for the Text Classification Task in the HoC dataset. 45
3.10 Experimental Results for Few-Shot Learning. Here, ’Precision’, ‘Recall’, ‘Readability-Controlled’, and ‘BERTScore’ are denoted by ’P‘, ’R’, RC’, and ‘B-S’, respectively whereas , ‘R-1’, ‘R-2’, and ‘R-L’ denote ‘ROUGE-1’, ‘ROUGE-2’, ‘ROUGE-L’, respectively. 46
3.11 Performance comparison fine-tuned smaller LLM (LLaMA-2-7b) with larger zero-shot and few-shot LLMs. 48

3.12 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. 50

A.1 Some examples of LLM generated responses in Relation Extraction datasets . . 83

A.2 Some examples of LLM generated responses in Dialogue Summarization datasets 84

A.3 Some examples of LLM generated responses in Question Summarization datasets 84

A.4 Some examples of LLM generated responses in Answer Summarization datasets 85

A.5 Some examples of LLM generated lay summaries in Biomedical Lay Summarization task in eLife dataset. 86

List of Abbreviations

AI	Artificial Intelligence
ALBERT	A Lite BERT
AnS	Answer Summarization
API	Application Programming Interface
B	Billion
BART	Bidirectional and Auto-Regressive Transformer
BC2GM	BioCreative II Gene Mention Recognition
BC4CHEMD	BioCreative IV Chemical compound and drug name recognition
BC5CDR	BioCreative V Chemical Disease Relation
BC5CDR-CHEM	BioCreative V Chemial Disease Relation-Chemical
BC5CDR-Disease	BioCreative V Chemial Disease Relation-Disease
BERT	Bidirectional Encoder Representations from Transformers
BioBART	Bidirectional and Auto-Regressive Transformers for the Biomedical Domain
BioBERT	Bidirectional Encoder Representations from Transformers for Biomedical Text Mining
BioGPT	Generative Pre-trained Transformer for Biomedical Text Generation and Mining
BioNLP	Biomedical Natural Language Processing
COMETA	Corpus of Online Medical EnTities
DDI	Drug-Drug Interaction

DTI	Drug-Target Interaction
ECG	Electrocardiogram
FLAN	Fine-tuned LAnguage Net
GPT	Generative Pre-training Transformer
HoC	Hallmarks of Cancer
LCS	Longest Common Subsequence
LLM	Large Language Models
LLaMA	Large Language Model Meta AI
MAS	Multi-Answer Summarization
MeQSum	Medical Question Summarization
NCBI	National Center for Biotechnology Information
NER	Named Entity Recognition
NLP	Natural Language Processing
PQA	PubMedQA
PQA-A	PQA-Artificial
PQA-L	PQA-Labeled
PQA-U	PQA-Unlabeled
PaLM	Pathways Language Model
PMC	PubMed Central
PLOS	Public Library of Science
QA	Question Answering
QS	Question Summarization
RLHF	Reinforcement learning from human feedback
RoBERTa	Robustly Optimized BERT-Pretraining Approach
ROUGE	Recall-Oriented Understudy for Gisting Evaluation

s800	Species-800
SOTA	State-of-the-art
T5	Text-To-Text Transfer Transformer

1 Introduction

The rapid growth of Large Language Models (LLMs) (Rogers et al., 2021) has led to dramatic progress in various domains, including biology (Kalyan et al., 2022). These powerful language models are trained on vast amounts of text data, allowing them to understand and generate natural language, identify patterns, and make predictions. This has opened up a new avenue of research and development, with applications ranging from protein design to drug discovery to personalized medicine (Zhang et al., 2024b; Madani et al., 2023).

In early years, specialized language models like BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) (Lee et al., 2020), BioBART (Bidirectional and Auto-Regressive Transformers for the Biomedical Domain) (Yuan et al., 2022a), BioGPT (Generative Pre-trained Transformer for Biomedical Text Generation and Mining) (Luo et al., 2022a), etc. have been proposed which demonstrated promising results in the biomedical domain. However, these models required task-specific finetuning on labeled datasets. Since fine-tuning requires a good amount of training examples (e.g., more than thousands of samples) to re-train a pre-trained language model on a specific task, this process is time-consuming due to the requirement of developing task-specific large annotated datasets, alongside requiring extensive machine learning expertise to train and evaluate these models. In contrast, zero-shot learning (Wang et al., 2019) enables a trained model to complete a task without the need for fine-tuning on task-specific datasets. The impressive zero-shot capabilities of

LLMs (Laskar et al., 2023a) make it possible to use them off-the-shelf to solve a wide range of tasks in various domains, without requiring any extensive skills in machine learning or computer programming to use them.

Nonetheless, in a sensitive discipline like biomedicine, it is important to ensure the usage of a system that is accurate and reliable. Since there are hundreds of LLMs that are currently available, it is important to choose the right LLM to use in this domain. While there are a large number of studies on various biomedical tasks (Liu et al., 2023d; Rahman et al., 2021; Morid et al., 2021; Cohen and Hersh, 2005; Wang et al., 2021), such as genomic sequence analysis (O'Brien et al., 2018; Ji et al., 2021b), disease diagnosis (Ali et al., 2021), drug discovery (Shaker et al., 2021; Martinelli, 2022; Pandiyan and Wang, 2022), cancer research (Nguyen et al., 2019), vaccine development (Soleymani et al., 2022), etc., it should be noted that the biomedical text processing task is closely related to all 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 regard, this thesis aims to conduct a comprehensive evaluation of popular LLMs in benchmark biomedical text processing tasks to identify their advantages and limitations. For this purpose, the existing LLM-based zero-shot state-of-the-art (SOTA) approaches are extensively evaluated. In the following, at first, the background of various core components of this research are presented. Afterward, the main hypotheses for this thesis are discussed. Then the research objectives are discussed. Finally, the main contributions to the thesis are demonstrated.

1.1 Background

Language modeling is the process of creating a model that can predict the likelihood of a sequence of words. This involves using statistical and probabilistic techniques to understand

and generate human language. In recent years, powerful language models have been built by pre-training them on massive amounts of text corpora. These language models are now the backbone of various biomedical Artificial Intelligence (AI) applications. In this section, at first various language models that are proposed in recent years in the biomedical domain are briefly reviewed, followed by a brief introduction to the LLMs that have been studied in this thesis. Afterward, the biomedical tasks that are studied in this thesis are demonstrated. Finally, various prompting techniques to interact with LLMs are discussed.

1.1.1 Language Models for The Biomedical Domain

In recent years, the effective utilization of neural networks based on the Transformer architecture (Vaswani et al., 2017) to build Natural Language Processing (NLP) models like BERT (Devlin et al., 2019) and GPT (Radford et al., 2019) have led to significant progress in the utilization of AI technologies. These neural models are first pre-trained on a vast amount of unlabeled text corpora in a self-supervised fashion for language modeling followed by fine-tuning on task-specific datasets to achieve SOTA results in various text processing tasks (Rogers et al., 2021). BERT leverages the encoder of the transformer architecture, this restricts models based on the BERT architecture to be only applicable to discriminative tasks like text classification. Meanwhile, GPT leverages the decoder of the transformer model which allows it to be used for text generation tasks. In addition to these models, sequence-to-sequence models like BART (Lewis et al., 2019) 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 (Yuan et al., 2022a). Figure 1.1 demonstrates various pre-training techniques depending on the model type.

It has been observed that further domain-specific pre-training of these models on the biomedical text corpora followed by fine-tuning on task-specific biomedical datasets (see Figure 1.2) have

helped these models to achieve SOTA performance in a variety of Biomedical NLP (BioNLP) tasks (Lee et al., 2020; Alsentzer et al., 2019; Beltagy et al., 2019; Gu et al., 2020; Peng et al., 2019; raj Kanakarajan et al., 2021; Gu et al., 2021). This has led to the development of various language models for the biomedical domain, such as BioBERT (Lee et al., 2020), Clinical-BERT (Alsentzer et al., 2019), BioBART (Yuan et al., 2022a), BioElectra (raj Kanakarajan et al., 2021), BioGPT (Luo et al., 2022a), ProteinBERT (Brandes et al., 2022), DNABERT (Ji et al., 2021b) 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 (Tajbakhsh et al., 2020). Nonetheless, a vast amount of unlabeled textual data is available in the field of biomedicine. While the scarcity of annotated datasets in this domain makes it difficult to build suitable models for biomedical tasks that lack large annotated datasets, 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 it was not exclusively trained on. Below, we demonstrate how larger language models address this problem.

1.1.2 Large Language Models

The task-specific fine-tuning of pre-trained language models on large annotated datasets led to impressive gains in performance. However, many tasks do not have large annotated datasets available, especially in specialized domains like biomedicine. This limitation of requiring task-specific large annotated datasets has led to the development of GPT-3 (Brown et al., 2020), a larger language model that is much larger in size than its predecessor GPT-2 (as well as other contemporary language models from the GPT-2 era, such as BERT), being trained on massive text corpora via leveraging extensive computing resources. This new generation of

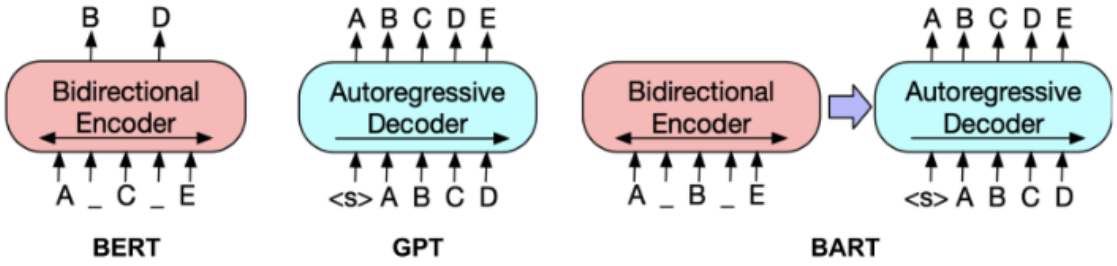


Figure 1.1: An overview of various pre-training techniques depending on the model type: on the left, (i) BERT for the encoder-based models to predict the masked words/tokens in a sequence (e.g., masked language modeling for text classification) based on the bidirectional context, then (ii) GPT for the decoder-based model to predict the next word in a sequence (e.g., auto-regressive language modeling for text generation) based on the context from the previous words, and finally (iii) BART for the encoder-decoder based model that utilizes both the encoder and the decoder of transformer with the encoders being used for text classification type objective while the decoders being used for text generation objectives. This figure is taken from the BART paper (Lewis et al., 2019).

large autoregressive[1] decoder-based language models like GPT-3 (Brown et al., 2020) have demonstrated impressive few-shot learning capability, instead of requiring any fine-tuning on task-specific large annotated datasets. With the success of GPT-3 in few-shot scenarios, a new variant of GPT-3 called the InstructGPT (in other words, GPT-3.5) model (Ouyang et al., 2022) has been proposed that demonstrates better capability at following instructions than the original GPT-3 model, resulting in an impressive zero-shot performance across a wide range of language processing tasks, including text classification, question answering, and text summarization (Qin et al., 2023; Bang et al., 2023; Yang et al., 2023). ChatGPT[2] 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[3] (Chowdhery et al., 2022; Anil et al., 2023; Singhal et al., 2023), Claude[4], LLaMA[5] (Touvron et al., 2023a,b) etc. have

[1] An autoregressive language model is a type of statistical model that is used to predict the next word in a text sequence based on the previous words in that sequence.

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

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

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

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

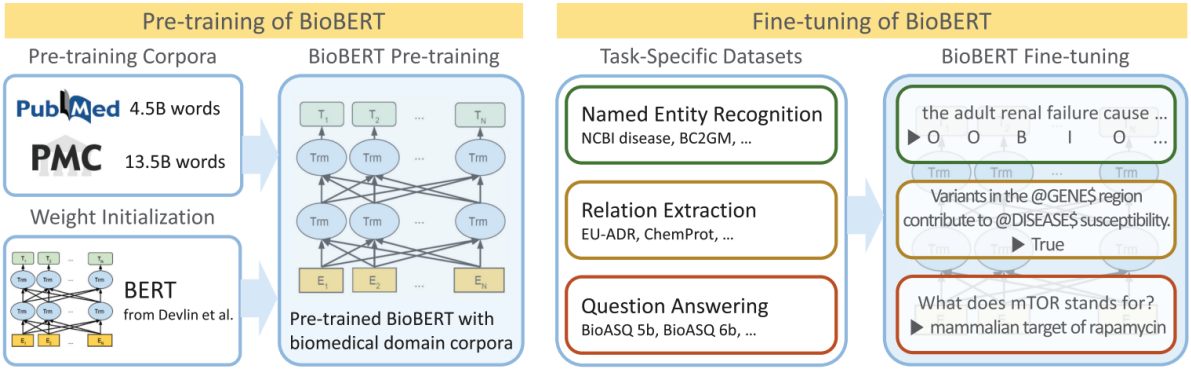


Figure 1.2: Overview of the continual pre-training of language models (e.g., BioBERT) for the adaptation to the biomedical domain followed by fine-tuning. This figure is also taken from the BioBERT paper (Lee et al., 2020).

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 (Kaelbling et al., 1996) from human feedback (RLHF) that leveraged both supervised learning and reinforcement learning techniques. While the supervised learning phase aimed to train the model to produce high-quality responses, 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 (see Figure 1.3) ranked the responses that the model had created in a previous conversation.

The capabilities of LLMs to also generate human-like responses alongside their zero-shot capabilities across various tasks have led to the development of LLM-powered automated tools. Systems built on top of LLM technologies can also be useful for researchers and practitioners in specialized domains like biology and medicine to find relevant information and extract insights from this vast corpus of unannotated biomedical data. However, despite being evaluated on various traditional NLP tasks, there is a lack of comprehensive studies that evaluate LLMs in the biomedical domain (Laskar et al., 2024).

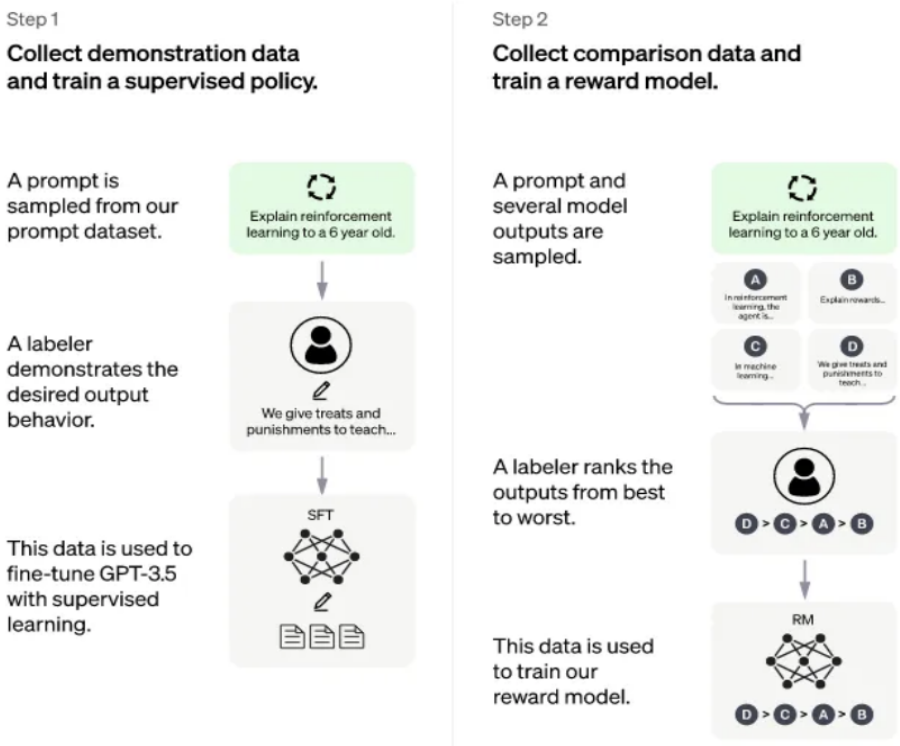


Figure 1.3: An overview of the overall training procedure of LLMs that combines supervised learning with reinforcement learning to train a reward model. This figure is taken from OpenAI (https://openai.com/index/chatgpt/). Last accessed June 1, 2024.

To this end, this research aims to evaluate LLMs across benchmark biomedical tasks. Below, these benchmark biomedical tasks have been demonstrated.

1.1.3 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:

$$T : X \rightarrow Y \tag{1.1}$$

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 research along with some examples are demonstrated.

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 (Yadav and Bethard, 2018). 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".

Biomedical relation extraction

The relation extraction task aims to extract relations between named entities in a given text (Zhong and Chen, 2021). 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”.

Biomedical entity linking

The entity linking task focuses on linking named entities in a text to their corresponding entries in a knowledge base (Laskar et al., 2022a,b). 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/drugin fo/taxol.

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.

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 (Abacha et al., 2019; He et al., 2020).

Biomedical text summarization

The main purpose of the text summarization task is to generate a short concise summary of the given document (El-Kassas et al., 2021; Laskar et al., 2022c). 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 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.

1.1.4 Prompting Large Language Models

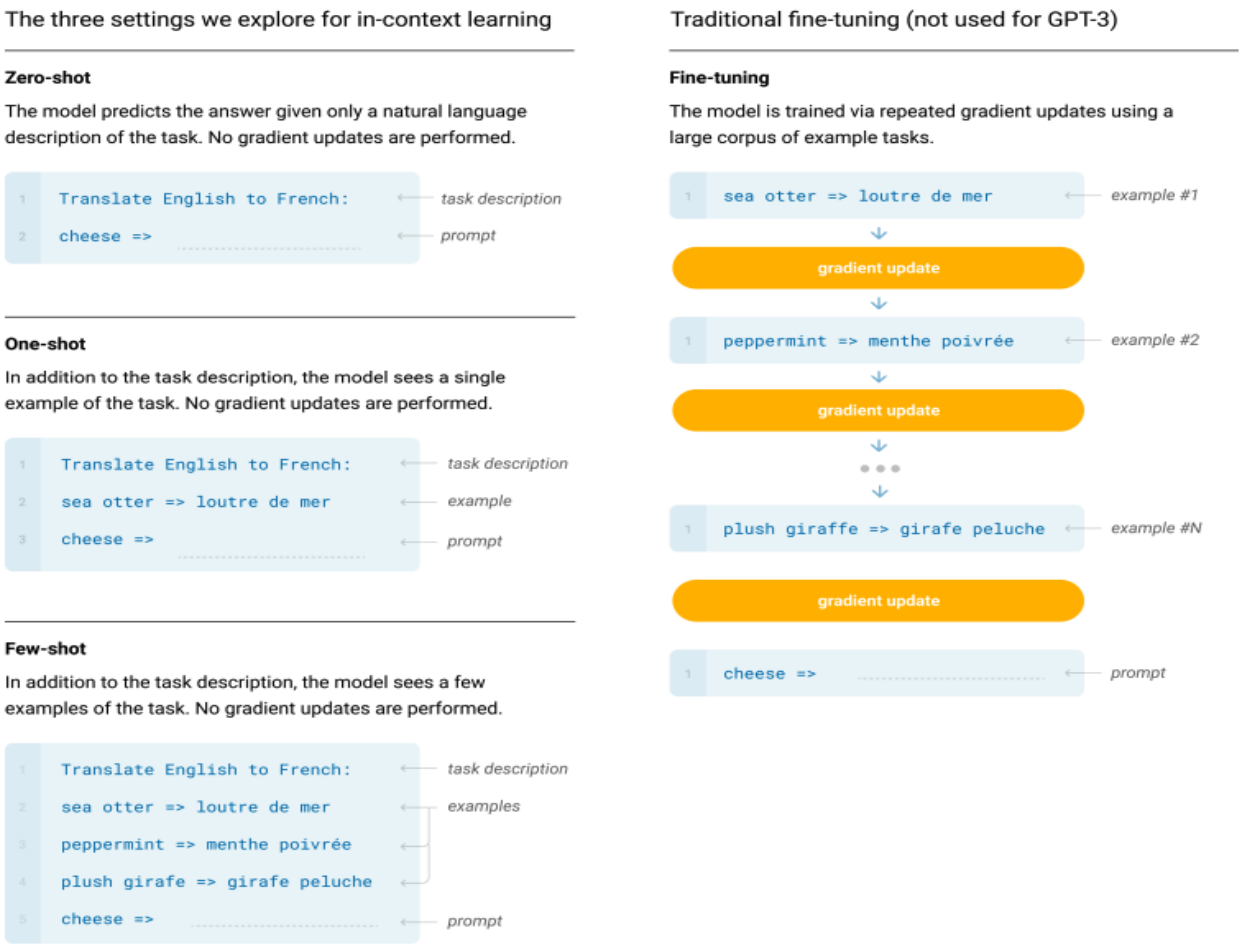


Figure 1.4: Overview of zero-shot, one-shot, few-shot, and traditional fine-tuning techniques. This figure is taken from the GPT-3 paper (Brown et al., 2020).

The utilization of LLMs in the biomedical domain would require a proper understanding of the complex linguistic characteristics of biomedical texts. In order to interact with LLMs, it is important to create prompts. A "prompt" refers to a piece of text or a set of instructions given

to an AI language model to guide its generation of text, which empowers it to produce better responses by grasping the provided context. Since LLMs are sensitive to prompts (Liu et al., 2023b; Jahan et al., 2023), it is important to design effective prompts to ensure the optimal utilization of LLMs. This can be achieved via prompt engineering. More specifically, prompt engineering is the process of designing prompts to effectively interact with and guide responses from LLMs. Thus, the goal of prompt engineering is to create a refined prompt that elicits the desired information or behavior from the model, enabling LLMs to even gain domain expertise and understand downstream tasks more effectively (Naveed et al., 2023).

LLMs can be prompted with both zero-shot and few-shot instructions. In zero-shot Learning, the model is required to identify and classify action categories that were not present during the training phase. Zero-shot utilization is becoming increasingly popular for several reasons, such as, eliminating the need for labeled examples, simplifying model development complexities without requiring fine-tuning on labeled data, and also assessing generalization to novel tasks. On the other hand, few-shot learning involves predicting new categories with only a limited number of labeled samples provided (Fink, 2004; Fei-Fei et al., 2006). It utilizes a small number of input-output examples closely associated with the test input as a prompt to improve the performance of LLMs. It's important to ensure that the selected input-output examples cover a wide range of diversity (Zhang et al., 2022). When there is just a single labeled example given, few-shot learning is called one-shot learning (Fei-Fei et al., 2006). Examples demonstrating zero-shot, one-shot, few-shot, and traditional fine-tuning are demonstrated in Figure 1.4.

Nonetheless, for both zero-shot and few-shot scenarios, the performance of LLMs may notably decline when no task demonstrations are provided. For task generalization, task demonstrations and intermediate reasoning steps like *chain-of-thought* learning (Wei et al., 2022) in the prompt play a crucial role because clear instructions are necessary for language models to comprehend

an unfamiliar task (Kojima et al., 2022; Wei et al., 2021; Lampinen et al., 2022).

With the increasing usage of LLMs in various domains like biology, healthcare, robotics, education, and scientific research, the necessity to use LLMs to perform certain tasks only based on fewer examples (or even no examples at all) without requiring any advanced machine learning techniques has become more important than ever. Thus, the effective construction of prompts based on prompt engineering is important to best utilize these LLMs in biomedical applications. In this regard, domain-specific knowledge in the biomedical domain could play a pivotal role in improving the performance of LLMs in biomedical tasks by developing useful and effective prompts.

1.2 Hypothesis

This thesis has three main hypotheses:

- With the release of many LLMs in recent years, it is difficult to determine which LLMs are suitable for what tasks. A comprehensive evaluation of LLMs across a wide range of biomedical tasks will help to identify the right LLM for specific biomedical tasks. Therefore, the hypothesis is that not a single LLM may outperform others in all biomedical tasks and so their performance may vary on different tasks.

- Domain-specific knowledge is required to understand the complex linguistic structure of biomedical texts for the development of task-specific prompts. Thus, the hypothesis here is that, the same LLM may perform differently based on prompt variations, while different LLMs may follow the instructions given in a prompt differently. Thus, a comprehensive evaluation of LLMs based on the constructed prompts will help to understand this hypothesis.

- Moreover, domain-specific knowledge is also required to ensure a reliable evaluation of LLMs in biomedical datasets and tasks. This will help to identify the limitations of such models in this domain to pave the way for the development of more powerful LLMs in the future. The hypothesis here is that many biomedical datasets will require the intervention of humans to ensure a reliable evaluation which may not be possible otherwise without extensive human evaluation.

1.3 Research Objectives

To achieve the goals of this research, the following objectives are pursued:

- **Benchmarking Large Language Models:** Evaluate the performance of popular LLMs, like ChatGPT, PaLM, Claude, and LLaMA on diverse biomedical tasks, including (i) biomedical Named Entity Recognition (NER), (ii) biomedical relation extraction, (iii) biomedical entity linking, (iv) biomedical text classification, (v) biomedical question answering, and (vi) biomedical text summarization. This will provide insights into the capabilities and limitations of different LLMs in various biomedical tasks which will be helpful to identify which LLMs are suitable for what task.

- **Effective Development of Textual Prompts:** Investigate the optimal approach to construct prompts for LLMs by leveraging domain-specific knowledge to improve their performance in the biomedical domain. This objective requires an understanding of the complex linguistic structure of biomedical texts for the development of effective prompts to ensure the most optimal utilization of LLMs.

- **Human Evaluation of Large Language Model Outputs:** Develop a reliable evaluation method for LLMs in the biomedical domain by utilizing domain-specific knowledge to

assess the complex linguistic structure of generated texts. This objective will address the challenges of the lack of domain experts in evaluating generative outputs in the biomedical domain.

2 Methodology

This section demonstrates how the prompts are designed for different datasets and tasks, describes the LLMs that have been studied in this thesis, and presents the evaluation pipeline. An overview of the methodology is also shown in Figure 2.1.

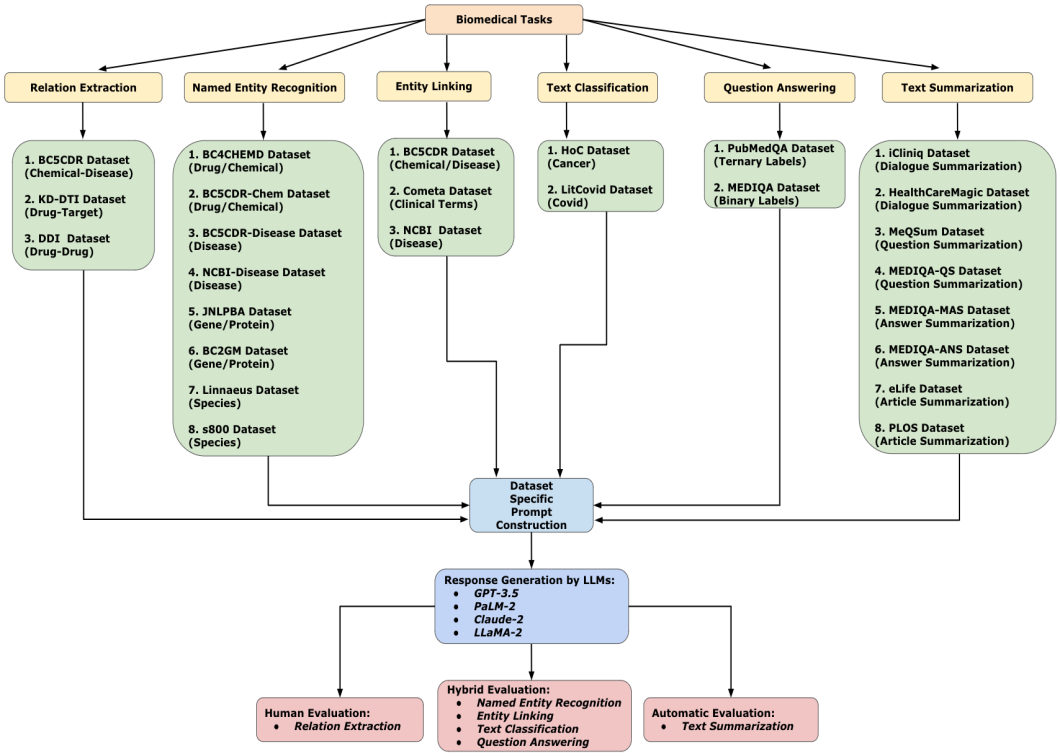


Figure 2.1: An overview of the methodology to evaluate 6 biomedical tasks across 26 datasets in this thesis. At first, the prompts are constructed for each dataset. Then, the responses are generated for each dataset using respective LLMs. Finally, depending on the task, various evaluation techniques are applied.

2.1 Datasets and Prompts

For a given test sample X in a dataset D, first a task instruction T is prepared. Then, the test sample X is concatenated 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 constructed for each task depending on the respective dataset has been demonstrated.

2.1.1 *NER*

The entities in the Named Entity Recognition (NER) task can be of various types, such as Person, Location, Organization, etc (Si et al., 2005). In the biomedical domain, the named entities can be Genes, Proteins, Diseases, etc (Huang et al., 2005).

In this thesis, prompts for NER are designed to identify the biomedical named entities in a given text in the BIO format. Note that the BIO format is a common tagging scheme used to label tokens in a text with their corresponding entity types. BIO stands for Beginning, Inside, Outside, and it helps to distinguish between the beginning of an entity, the continuation of an entity, and tokens that are not part of any entity.

B - (Beginning): This tag is used to indicate the beginning of an entity. For instance, in the phrase "John Doe", "John" would be tagged as "B-PER" (Beginning of a Person).

I - (Inside): This tag indicates that the token is inside an entity that started with the B-tag. For example, in "John Doe", "Doe" would be tagged as "I-PER" (Inside the Person type named entity).

O - (Outside): This tag is used for tokens that are not part of any entity. For instance, in the sentence "John Doe is a software engineer", words like "is", "a", "software", and "engineer" would be tagged as "O".

An example sentence with BIO tagging might look like this:

"John B-PER Doe I-PER is O a O software O engineer O . O"

To construct the prompt for NER, the description of the BIO format is also added along with the task instructions. For NER, the BC2GM (Smith et al., 2008) and the JNLPBA (Collier and Kim, 2004) datasets are used for gene/protein entity recognition, BC4CHEMD (Krallinger et al., 2015) and BC5CDR-CHEM (Li et al., 2016) datasets are used for drug/chemical entity recognition, BC5CDR-Disease (Li et al., 2016) and NCBI-Disease (Dougan et al., 2014) are used for disease type entity recognition, LINNAEUS (Gerner et al., 2010) and s800 (Pafilis et al., 2013) are used for species type entity recognition. The prompts for this task are shown in Table 2.1.

Table 2.1: Sample Prompts in Different Named Entity Recognition (NER) Datasets.

Dataset	**Type**	**Data Split (Train / Valid / Test)**	**Prompt**
BC2GM	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	Drug / Chemical	30682 / 30639 / 26364	
BC5CDR-CHEM	Drug / Chemical	4560 / 4581 / 4797	
BC5CDR-Disease	Disease	4560 / 4581 / 4797	
JNLPBA	Gene / Protein	14690 / 3856 / 3856	
LINNAEUS	Species	11935 / 4078 / 7142	
NCBI-Disease	Disease	5424 / 923 / 940	
s800	Species	5733 / 830 / 1630	

2.1.2 *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, the prompts are constructed for chemical-disease-relation in the BC5CDR dataset (Li et al., 2016), drug-target-interaction in the KD-DTI dataset (Hou et al., 2022), and drug–drug-interaction in the DDI dataset (Herrero-Zazo et al., 2013). In the BC5CDR dataset, the LLM is prompted to identify the drug and drug-induced side effects. In KD-DTI, the LLM is prompted to identify the drug-target interactions alongside the interaction types. In the DDI dataset, the LLM is again asked to identify the drug-drug interactions alongside the interaction types. Since the possible interaction types in DDI is only 4 (whereas in KD-DTI there are more than 20 interaction types), the definition is also added for each interaction type for DDI. The prompts used for these datasets are demonstrated in Table 2.2.

Table 2.2: Sample Prompts in Different Relation Extraction Datasets.

Dataset	**Type**	**Data Split (Train / Valid / Test)**	**Prompt**
BC5CDR	Chemical-Disease	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	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	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]

2.1.3 *Entity linking*

To identify whether LLMs can link named entities to their correct definitions based on their pre-training knowledge, the work of Yuan et al. (2022b) has been followed for the generative entity linking task by asking LLMs to identify the correct concept names for the named entities. For evaluation, the BC5CDR (Li et al., 2016) dataset for the entity linking of disease/chemical type named entities, the NCBI (Dougan et al., 2014) dataset to link diseases, and the COMETA (Basaldella et al., 2020) dataset to link clinical terms have been used. The sample prompts for this task are shown in Table 2.3.

Table 2.3: Sample Prompts in Different Entity Linking Datasets.

Dataset	**Type**	**Data Split (Train / Valid / Test)**	**Prompt**
BC5CDR	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	Clinical Terms	13489 / 2176 / 4350	
NCBI	Disease	5784 / 787 / 960	

2.1.4 *Question answering*

For the question-answering task, the performance of LLMs has been evaluated on the PubMedQA dataset (Jin et al., 2019) and the MEDIQA-2019 dataset (Abacha et al., 2019). 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 (Laskar et al., 2020). In the MediQA-2019 dataset, there are 4 different labels, where 4 indicate *excellent answers*, 3 indicate *correct but incomplete answers*, 2 indicate *related answers*, and 1 indicates *incorrect answers*. In this evaluation, the work of Xu et al. (2019) is followed and the labels 3 and 4 are considered as Relevant

while 1 and 2 are considered as Not Relevant. The prompts for this task are shown in Table 2.4.

Table 2.4: 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]

2.1.5 *Text classification*

The goal of this task is to classify the type of the given text. In this thesis, two datasets are used: the HoC (the Hallmarks of Cancer corpus) dataset (Baker et al., 2016), and the LitCovid dataset (Chen et al., 2021). 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[1]. 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*. The prompts for these text classification datasets are shown in Table 2.5.

2.1.6 *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 (Abacha

[1] https://en.wikipedia.org/wiki/The_Hallmarks_of_Cancer

Table 2.5: 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]

and Demner-Fushman, 2019) and MEDIQA-QS (Abacha et al., 2021) datasets), medical answer summarization (MEDIQA-ANS (Savery et al., 2020) and MEDIQA-MAS (Abacha et al., 2021) datasets), and doctor–patient dialogue summarization (iCliniq and HealthCareMagic datasets (Zeng et al., 2020; Mrini et al., 2021)) to generate short queries for healthcare forums describing patient's medical conditions. In addition, various datasets are used for biomedical literature summarization (Luo et al., 2022b; Goldsack et al., 2022), such as the Biomedical Text Lay Summarization shared task 2023 (BioLaySumm-2023) datasets (Goldsack et al., 2023). For BioLaySumm-2023, since the gold reference summaries of the test sets are not publicly available as of the writing of this thesis, the respective validation sets are used for evaluation. The sample prompts in the summarization datasets are shown in Table 2.6.

Table 2.6: Sample Prompts in Different Text Summarization tasks.

Dataset	**Type**	**Data Split (Train / Valid / Test)**	**Prompt**
iCliniq	Dialog 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]
HealthCareMagic	Dialog 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 (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 (Abstract)	24773 / 1376 / 142	Write down the abstract of the following biomedical article: [ARTICLE]

2.2 Models

In the following, the 4 popular LLMs that are evaluated in benchmark biomedical datasets and tasks in this thesis are presented.

2.2.1 *GPT-3.5*

GPT-3.5 is an auto-regressive language model based on the transformer (Vaswani et al., 2017) 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 (OpenAI, 2023). 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 publicly available text, including articles, books, and websites. Additionally, OpenAI collected data from GPT-3 users to train and fine-tune the model further (Qin et al., 2023; OpenAI, 2023). In this work, the OpenAI API is used for the gpt-3.5-turbo-0613[2] model for GPT-3.5.

2.2.2 *PaLM-2*

PaLM-2 (Anil et al., 2023) 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, 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, making it smaller than its predecessor PaLM-1, while achieving better performance. This helps PaLM-2 to achieve faster inference with fewer parameters, alongside ensuring lower serving costs. In this work, the text-bison@001[3] model in Google's Vertex AI[4] API is used for PaLM-2.

[2] https://platform.openai.com/docs/models/gpt-3-5-turbo
[3] https://cloud.google.com/vertex-ai/generative-ai/docs/model-reference/text
[4] https://cloud.google.com/vertex-ai

2.2.3 *Claude-2*

Claude-2 is also a general-purpose LLM based on the transformer architecture. It was developed by Anthropic[5] 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 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, Claude-2 is utilized in this work via leveraging Anthropic's claude-2 API.

2.2.4 *LLaMA-2*

LLaMA-2 (Touvron et al., 2023b) is a recently proposed LLM by Meta. One major advantage of LLaMA-2 over the previously mentioned LLMs is that it is open-sourced. While another open-sourced version of LLaMA: the LLaMA-1 (Touvron et al., 2023a) 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 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 finetuning, 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, while the 7B model requires less computational resources, it achieves poorer performance in comparison to the 13B and 70B

[5] `https://www.anthropic.com/claude`

models. Considering the performance and cost trade-off, the LLaMA-2-13B[6] model is used in this work.

2.3 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 research, to ensure high-quality evaluation, the work of Laskar et al. (Laskar et al., 2023a) is followed where they design different settings for the evaluation of LLMs for different tasks, as demonstrated below:

Automatic Evaluation: Tasks that depend on metrics like string matching (e.g., ROUGE (Lin, 2004a)) or contextual similarity (e.g., BERTScore (Zhang et al., 2019)) do not require the parsing of the LLM-generated responses and so a direct comparison between the full response generated by LLMs against the gold reference can be applied.

Hybrid (Human + Automatic) Evaluation: Since LLMs generate human-like free-form text, in discriminative tasks, it is not trivial to directly compare LLM-generated responses with the gold label. This is because the responses generated by LLMs in such tasks may contain texts additional to what is required for a certain task, alongside the presence of synonyms or abbreviations. Thus, for such tasks, a programming script is usually written to extract the required answer from the LLM-generated response. Nonetheless, LLMs may not follow a specific for-

[6] https://huggingface.co/meta-llama/Llama-2-13b-chat-hf

mat for response generation and so programming scripts may not always be able to extract the required answers from a wide variety of responses. To address this problem, the intervention of humans is required. More specifically, this is done by first applying programming scripts on the dataset to parse the required answer from the LLM-generated response, followed by utilizing human intervention if solely depending on the programming script cannot extract the results in the expected format. This hybrid evaluation technique is also used by Laskar et al. (2023a) to evaluate ChatGPT.

Human Evaluation: As demonstrated in early research (Laskar et al., 2023a), some discriminative tasks require the full intervention of humans for evaluation. This is due to the descriptive nature of LLM-generated responses, and the presence of a higher number of synonyms and abbreviations, making the programming scripts fail in most cases to extract the required answer.

In this thesis, the evaluation methodologies described above are followed for the evaluation of LLMs in various biomedical tasks.

For discriminative tasks, where parsing of the results from the generated response is required for evaluation, the work of Laskar et al. (Laskar et al., 2023a) is followed and a programming script is designed 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, the performances are evaluated 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 (Laskar

et al., 2023a; Jahan et al., 2023).

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

2.4 Evaluation Metrics

Different evaluation metrics for different tasks are used to ensure a fair comparison of different LLMs with prior SOTA 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 NER tasks, Precision, Recall, and F1 metrics are used, while for entity linking, only Recall metric is used. For summarization, the ROUGE (Lin, 2004a) and the BERTScore (Zhang et al., 2019) metrics are used. For question answering and text classification, metrics like Accuracy and F1 are used. Below each evaluation metric is described for different tasks:

Precision: Precision is the proportion of correctly identified positive cases out of all cases predicted as positive. The formula for precision is:

$$Precision = \frac{TP}{TP + FP}$$

Here,

TP (True Positives) are the cases where the model correctly predicts the positive class.

FP (False Positives) are the cases where the model incorrectly predicts the positive class.

Recall: Recall is the proportion of correctly identified positive cases out of all actual positive cases. The formula for recall is:

$$Recall = \frac{TP}{TP + FN}$$

Here,

TP (True Positives) are the cases where the model correctly predicts the positive class.

FN (False Negative) are the cases where the model inaccurately predicts the positive class as negative.

F1 Score: The F1 score is the harmonic mean of precision and recall. This score is a balance of both precision and recall. The formula for the F1 score is:

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

ROUGE: ROUGE (Lin, 2004a) is a group of evaluation metrics used to assess the quality of generated text, particularly in tasks like summarization and translation. It examines the overlap of n-grams (sequences of n words) between the generated text and reference summaries.

ROUGE-N: Measures the overlap of n-grams between the generated text and the reference summaries. ROUGE-N is typically computed for different values of n, such as ROUGE-1 (unigram overlap), ROUGE-2 (bigram overlap), and so on.

ROUGE-L: Computes the longest common subsequence (LCS) between the generated text and the reference summaries. It measures the longest sequence of words that appear in both the generated text and the reference summaries.

BERTScore: BERTScore (Zhang et al., 2019) is a metric that evaluates the quality of the generated text by comparing the contextual embeddings of words in the generated text with those in the reference. It computes similarity scores for each word in the generated text with respect to the reference and then aggregates them to provide an overall quality score. BERTScore is

effective in capturing semantic similarity in the generated text.

Accuracy: Accuracy measures the overall correctness of the model's predictions across all classes. It calculates the proportion of correctly predicted cases (both positive and negative) out of all cases. The formula for accuracy is:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Here,

TN (True Negatives) are the cases where the model correctly predicts the negative class.

FP (False Positives) are the cases where the model incorrectly predicts the positive class.

FN (False Negatives) are the cases where the model fails to predict the positive class when it is actually positive.

2.5 Baselines

To compare the performance of the zero-shot LLMs, models that achieve SOTA performance in various biomedical text processing tasks are used as the baselines. These baseline models are described below:

BioGPT: The backbone of BioGPT (Luo et al., 2022a) is GPT-2 (Radford et al., 2019), which is a decoder of the transformer (Vaswani et al., 2017). The BioGPT model was trained over PubMed titles and abstracts via leveraging the standard language modeling task. The fine-tuned BioGPT models are used as the baseline for all datasets in the relation extraction task, HoC dataset in the text classification task, and the PubMedQA[7] dataset for the question-answering tasks.

BioBART: It is a sequence-to-sequence model based on the BART (Lewis et al., 2019) archi-

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

tecture where the pre-training process involves reconstructing corrupted input sequences. The main difference between BioBART (Yuan et al., 2022a) and BART is that the former was pre-trained over PubMed abstracts to make it suitable for the biomedical domain tasks. In this work the finetuned BioBART model is used as the baseline in all the entity linking datasets and the following biomedical summarization tasks: Dialogue Summarization, Question Summarization, and Answer Summarization.

BioBERT: It is a domain-specific language representation model (Lee et al., 2020) based on the BERT (Devlin et al., 2019) architecture that was additionally pretrained on large-scale biomedical corpora (PubMed and PMC abstracts). The fine-tuned BioBERT model achieved SOTA performance across different biomedical NER datasets and so it is used as the baseline for all NER datasets in this paper. In addition, it is used as the baseline in the LitCovid dataset for text classification.

ALBERT: The ALBERT (Lan et al., 2019) model is a variant of the BERT (Devlin et al., 2019) language model which requires lower memory consumption. He at al., (He et al., 2020) extends its training mechanism by additionally training ALBERT on 14K biomedical texts in a question-answering fashion via infusing disease knowledge which led to the SOTA performance in the MediQA-2019 dataset. The LLMs are compared with this disease knowledge infused version of the ALBERT model in this work.

FLAN-T5-XL: FLAN-T5 (Chung et al., 2022) is an extension of the T5 (Raffel et al., 2020) model. The T5 model is an encoder-decoder model that treats each task as a sequence-to-sequence problem. While the architecture of FLAN-T5 is similar to the original T5 model, it leverages instruction fine-tuning instead of traditional fine-tuning. The FLAN-T5-XL achieves SOTA performance in the Biomedical Lay Summarization task. Therefore, it is used as the baseline in the eLife and the PLOS datasets to compare LLMs in biomedical lay summarization.

PRIMERA: It is a pre-trained model (Xiao et al., 2022) designed to enhance multi-document

summarization. It proposes a new pre-training strategy for multi-document summarization by leveraging the longformer-encoder–decoder (Beltagy et al., 2020) for pre-training. In this work, the fine-tuned PRIMERA model is used as the baseline in the readability-controlled[8] summarization task since it is the current SOTA in this task.

[8]In the readibility-controlled summarization, the summary of the biomedical article is required to be generated based on only the given biomedical article text without the abstract of the article.

3 Results and Discussion

3.1 Performance Comparisons

In this section, the results for LLMs in various tasks are presented and compared with the current SOTA models. At first, the results in the relation extraction task are presented where the human evaluation is utilized. Then, the findings in text classification, question answering, entity linking, and NER datasets are demonstrated where hybrid evaluation is conducted. Finally, the findings in the summarization datasets are presented where automatic evaluation is utilized.

3.1.1 *Relation extraction*

The performance of LLMs is compared with the current SOTA fine-tuned BioGPT (Luo et al., 2022a) 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 3.1, it is found 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 SOTA 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 SOTA 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 SOTA 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 12K training samples, most zero-shot LLMs still achieve comparable performance, with PaLM-2 slightly outperforming the SOTA result. More interestingly, while other LLMs achieve quite poor precision scores in the KD-DTI dataset, PaLM-2 even outperforms the current SOTA 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**.

Table 3.1: Performance on Relation Extraction datasets. All SOTA results are taken from the BioGPT (Luo et al., 2022a) 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	49.52	43.25	46.17	40.00	39.72	38.42	41.70	44.75	40.76

3.1.2 *Text classification*

In terms of text classification (see Table 3.2), the LLM generated responses are evaluated based on **Hybrid Evaluation**. In comparison to the current SOTA models fine-tuned on the respective datasets (BioGPT (Luo et al., 2022a) in HoC and BioBERT (Lee et al., 2020) in LitCovid), it is evident that the zero-shot LLMs perform very poorly in comparison to the SOTA 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 $p \leq .05$.

Table 3.2: Performance on Text Classification, Question Answering, and Entity Linking datasets. The SOTA results for HoC and PubMedQA are taken from the BioGPT (Luo et al., 2022a) model, while for LitCovid and MediQA-2019 datasets, the SOTA results are taken from Gutiérrez et al. (2020) and He et al. (2020), respectively. Note that all SOTA results for Entity Linking are taken from the BioBART (Yuan et al., 2022a) model.

Model	Text Classification		Question Answering		Entity Linking		
	HoC	LitCovid	PubMedQA	MediQA-2019	BC5CDR	Cometa	NCBI
	F1	F1	Accuracy	Accuracy	Recall	Recall	Recall
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	**85.12**	**86.20**	**78.20**	**79.49**	**93.26**	**81.77**	**89.90**

3.1.3 *Question answering*

For question answering, the performance is evaluated based on **Hybrid Evaluation** on two datasets (see Table 3.2).

In terms of the question-answering task in the PubMedQA dataset, it is found that the

performance of all LLMs is much lower than the current SOTA BioGPT model. It should be noted that the BioGPT (Luo et al., 2022a) model which achieves the SOTA 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), it is found 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, the confusion matrix is presented here using a heatmap based on the prediction made by different LLMs in Figure 3.1. From the heatmap, it shows 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 predicting the "yes" type label instead of predicting the expected "no" label, leading to an overall poor accuracy.

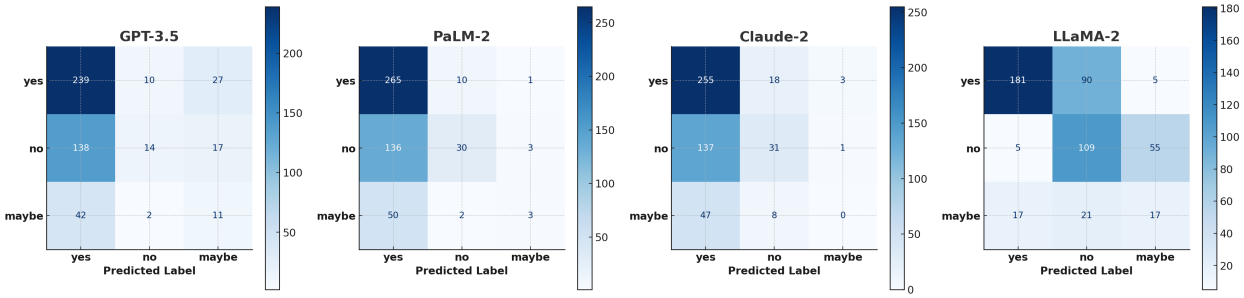


Figure 3.1: Confusion Matrix for different models in the PubMedQA dataset.

In terms of the question-answering task in the MediQA-2019 dataset, it is found 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 ac-

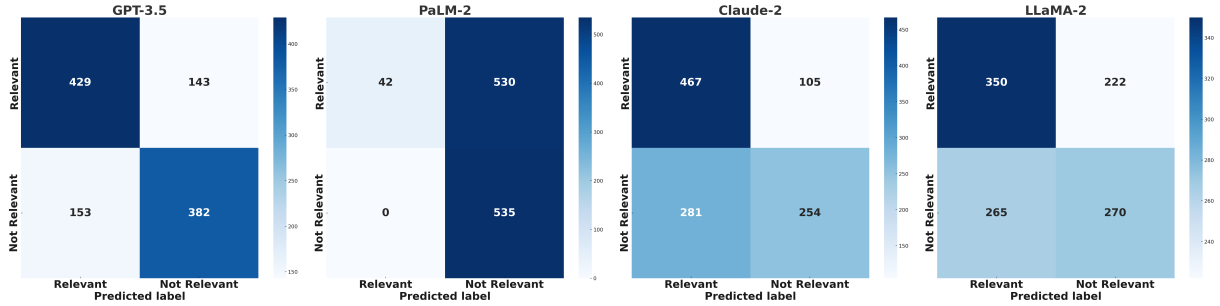


Figure 3.2: Confusion Matrix for different models in the MediQA-2019 dataset.

curacy (73.26). The performance of GPT-3.5 is comparable to the current SOTA accuracy of 79.49 (He et al., 2020) by the ALBERT model (Lan et al., 2019) which was additionally trained in question-answering style on 14K 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, the confusion matrix is showed in Figure 3.2 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. Meanwhile, 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.**

3.1.4 *Entity linking*

All the entity linking datasets are evaluated based on the **Hybrid Evaluation** technique. For entity linking, it is found from Table 3.2 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 $p \leq .05$.

3.1.5 *NER*

Similar to Entity Linking, **Hybrid Evaluation** is conducted for NER. It can be found from Table 3.3 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 (Laskar et al., 2023a). These findings give a strong indication that generative LLMs need further improvement on sequence labeling tasks like NER using the traditional BIO formatting.

Table 3.3: Performance on Named Entity Recognition datasets. SOTA results are from the BioBERT (Lee et al., 2020) model. Here, 'Precision' and 'Recall' are denoted by 'P' and 'R', respectively.

Dataset	**Model**														
	GPT-3.5			**PaLM-2**			**Claude-2**			**LLaMA-2-13b**			**State-of-the-Art**		
	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**

3.1.6 *Summarization*

The results for the following summarization datasets: dialog summarization, question summarization, and answer summarization are presented in Table 3.4 and compared with BioBART (Yuan et al., 2022a). For evaluation, the following two **Automatic Evaluation** metrics are used: (i) the widely used ROUGE (Lin, 2004b) metric, and (ii) the BERTScore (Zhang et al., 2019) metric. For BERTScore, the RoBERTa-Large (Liu et al., 2019) model is used for implementation. For all LLMs, the input context length of 2000 words has been used.

Table 3.4: 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. The SOTA (State-of-the-Art) results are taken from the BioBART (Yuan et al., 2022a) 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

It is observed 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 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 (Yuan et al., 2022a). This further confirms that zero-shot LLMs are more useful than domain-specific fine-tuned models in biomedical datasets that lack large training data.

Table 3.5: Performance on the Biomedical Lay Summarization task. SOTA results are from Sim et al. (2023).

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 3.6: Performance on Readability Controlled Summarization in the PLOS dataset. SOTA results are from Chen et al. (2023a).

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

Next, the experimental results on the biomedical lay summarization task and the readability-controlled summarization task are presented in Table 3.5 and Table 3.6, respectively.

For the biomedical lay summarization task, both abstract and article are combined together and given as input to the models till the concatenated text reaches the maximum context length. For this task, the performance of the LLMs are combined 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 SOTA 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 SOTA result. This shows a great discrepancy between the lexical matching-based 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 an article is only given as input without the abstract, as required by the task. It is found 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, it is found 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 (Laskar et al., 2023a).

Since the whole document cannot be given as input at once to these LLMs except Claude-2, the performances are investigated using the 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. Moreover, LLaMA-2 has a maximum context length of 4000 tokens (approximately 3000 words). Therefore, LLaMA-2 is excluded from this study. The results for both tasks, the biomedical lay summarization and the readability-controlled summarization can be found in Table 3.7 and Table 3.8, respectively. The 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. (Liu et al., 2023a), 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

Table 3.7: Performance of different LLMs on Biomedical Lay Summarization datasets based on various input lengths. Here, 'R-1', 'R-2', 'R-L' and 'B-S' denote 'ROUGE-1', 'ROUGE-2', 'ROUGE-L', and 'BERTScore', respectively.

Model	Length	Dataset							
		eLife				PLOS			
		R-1	R-2	R-L	B-S	R-1	R-2	R-L	B-S
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	10000	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 3.8: Performance of different LLMs on Readability Controlled Summarization in the PLOS dataset based on various input lengths. Here, 'R-1', 'R-2', 'R-L' and 'B-S' denote 'ROUGE-1', 'ROUGE-2', 'ROUGE-L', and 'BERTScore', respectively.

Model	Length	Summarization Type							
		Abstract				Lay Summarization			
		R-1	R-2	R-L	B-S	R-1	R-2	R-L	B-S
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	10000	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

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 (Laskar et al., 2023b).

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.

3.2 Analyses

In this section, the performance of LLMs is further analyzed based on (i) variations in prompts, (ii) few-shot learning, (iii) fine-tuning, (iv) analyzing the performance of LLMs based on the possibility of data contamination, (v) the average performance per task, and finally (vi) the computing cost associated with each LLM. Below, the findings based on these analyses are demonstrated.

3.2.1 *Effects of prompt variation*

The effects of prompt engineering 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 3.9).

ii. Prompting without mentioning the name of any HoC classes, drops F1 to 38.20 (see Row 2 in Table 3.9).

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 3.9).

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

3.2.2 *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, this section

Table 3.9: Effects of Prompt Variations in GPT-3.5 for the Text 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

demonstrates the effect of utilizing 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 research, 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 to address the context length issue. In the prompt, the few-shot examples are first included, followed by the task descriptions. The results from the few-shot experiments across all datasets are shown in

Table 3.10.

Table 3.10: Experimental Results for Few-Shot Learning. Here, 'Precision', 'Recall', 'Readability-Controlled', and 'BERTScore' are denoted by 'P', 'R', RC', and 'B-S', respectively whereas , 'R-1', 'R-2', and 'R-L' denote 'ROUGE-1', 'ROUGE-2', 'ROUGE-L', respectively.

Dataset	**Claude-2 (0-Shot)**			**Claude-2 (1-Shot)**			**Claude-2 (3-Shot)**			**SOTA**		
NER	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**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	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**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		**R**			**R**			**R**			**R**	
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**		

Though few-shot learning usually leads to improvements in performance, based on the results presented in Table 3.10, it is evident that few-shot learning does not necessarily lead to gain in performance in each task. More specifically, in many tasks, few-shot learning is found to be ineffective. Nonetheless, similar findings are also observed in recent research. For

instance, Ye et al. (2023) demonstrate that in many language processing tasks, few-shot learning using LLMs achieves much poorer results in comparison to zero-shot learning. From the results presented in Table 3.10, it is also found 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, 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 these findings are consistent with Ye et al. (Ye et al., 2023) 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.

3.2.3 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, the performance of the LLMs used in this thesis are compared with a comparatively smaller open-source LLM: the LLaMA-2-7b-Chat[1] model. This model was fine-tuned by Jahan et al. (2024) only on some biomedical datasets that have smaller training sets. Jahan et al. (2024) fine-tuned the LLaMA-2-7b-Chat model on the PubMedQA dataset for question-answering, the MeQSum dataset for summarization, and the DDI and BC5CDR datasets for relation extraction. For fine-tuning, 450 training samples in the PubMedQA dataset for question-answering, 500 training samples for summarization in the MeQSum dataset, and 664 training samples in the

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

Table 3.11: Performance comparison fine-tuned smaller LLM (LLaMA-2-7b) with larger zero-shot and few-shot LLMs.

Model	Relation Extraction Task						QA Task	Summarization Task			
	BC5CDR			DDI			PubMedQA	MeQSum			
	Precision	Recall	F1	Precision	Recall	F1	Accuracy	ROUGE-1	ROUGE-2	ROUGE-L	BertScore
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**

BC5CDR dataset while 500 training samples in the DDI dataset have been used for relation extraction, respectively. The responses generated by the LLaMA-2-7b-Chat model are then evaluated using the hybrid evaluation approach for PubMedQA, human evaluation for relation extraction, and automatic evaluation for summarization. The performance comparison between the zero-shot and few-shot larger LLMs with the fine-tuned LLaMA-2-7b is shown in Table 3.11. From Table 3.11, 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). Moreover, it achieves almost similar performance in comparison to the SOTA in the PubMedQA dataset for the question-answering task (even though LLaMA-2-7b was only trained on 500 samples, the current SOTA BioGPT (Luo et al., 2022a) model was trained on 270K samples). 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[2]. Thus, fine-tuning open-source LLMs like LLaMA-2 on more biomedical tasks could be an interesting future research direction.

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

3.2.4 *Data contamination detection analysis*

Recent research has demonstrated that there is a strong possibility of data contamination in many datasets that are currently used for the evaluation of LLMs (Xu et al., 2024; Ravaut et al., 2024; Li and Flanigan, 2023). In this section, the possibility of contamination in the datasets used for evaluation in this research have been studied. For this purpose, the work of Li and Flanigan (2023) is followed to analyze the possibility of the contamination of the datasets that has been studied in this research to evaluate various LLMs. For this purpose, the following has been done:

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 have been evaluated in this research.

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.

From the results for the data contamination detection analysis presented in Table 3.12, it can be inferred that in the NER datasets, none of the LLMs could extract the task examples. This could be due to the fact that while conducting the experiments for NER in this research, LLMs were asked to determine the NER tag for each token based on the 'BIO' format. Meanwhile, LLMs could potentially be pre-trained differently for the NER task. In this analysis, it is also found 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 'BIO format' in the pre-training data could probably be the reason behind LLMs performing very poorly in all NER datasets. A similar trend is also

Table 3.12: 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

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 technique shows 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 SOTA results. While for HoC, it is expected since all LLMs perform much poorer than the SOTA results. 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, it is found that only the PaLM-2 model has the possibility of data contamination. Nonetheless, 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). It is also found 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 SOTA models (with the results being much lower in comparison to the reported SOTA results). It should also be pointed out that the membership inference shows no possibility of contamination in datasets that are released in 2023.

3.2.5 *Average performance per task*

To provide further insights, the average performance of different LLMs across each task is presented in Figure 3.3 which demonstrates that:

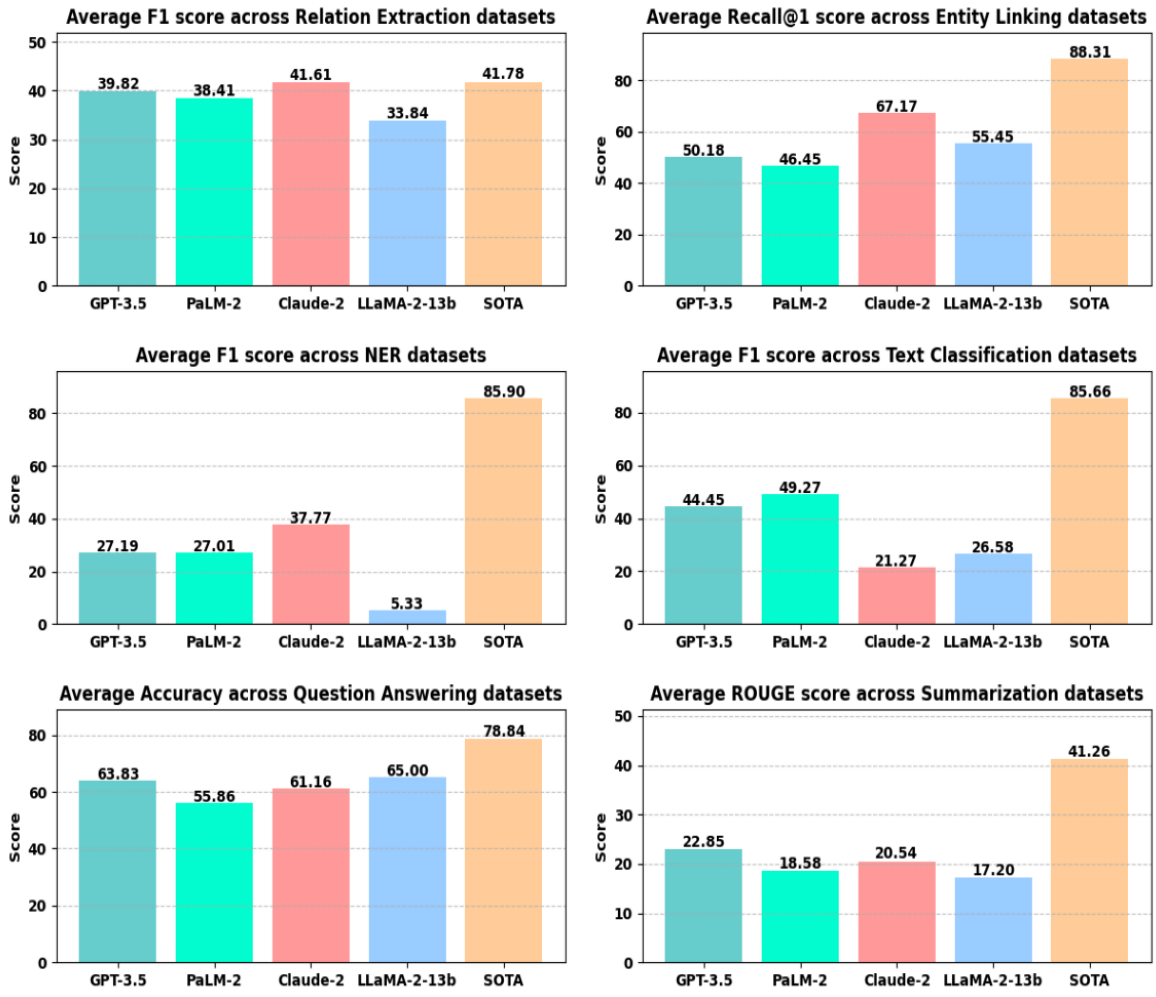


Figure 3.3: Average score by different zero-shot LLMs across each task.

- In Relation Extraction, the performance of most LLMs (e.g., GPT-3.5, PaLM-2, Claude-2) is comparable to the current SOTA results. However, LLaMA-2 lags quite behind.
- In Entity Linking and NER, the performance of all LLMs is far behind the current SOTA.

- In all entity related tasks (e.g., Relation Extraction, NER, Entity Linking), Claude-2 is found to be the best-performing LLM. Thus, the experimental results imply that Claude-2 could be the best LLM to use off-the-shelf for entity-related biomedical tasks.

- Nonetheless, in text classification, Claude-2 is found to be the worst-performing LLM. Similar to Entity Linking and NER, the performance of LLMs is also quite behind the current SOTA results.

- In non-entity related tasks, the experimental results demonstrate that PaLM-2 is the best LLM in text classification, GPT-3.5 is the best LLM in Summarization, while LLaMA-2- is the best LLM in Question Answering.

3.2.6 Cost Analysis

As of November 1, 2023, when all the experiments using GPT-3.5 and PaLM-2 were conducted for this thesis, the GPT-3.5-Turbo-0613 model that has been used costs 0.0015$ per 1K input tokens and 0.002$ per 1K output tokens. Meanwhile, for PaLM-2, the pricing[3] for both input and output is 0.0010$ per 1000 characters. Approximately, 1 token is considered as 4 characters (Laskar et al., 2023b). Thus, the cost for PaLM-2 is 0.0040$ per 1K tokens, making it about 2.5 times more costly than GPT-3.5. Meanwhile, for Claude2[4], the cost per 1 million input tokens is 8$, while for output tokens is 24$. This makes it 0.008$ per 1K input tokens while 0.024$ per 1K output tokens on average, making it quite costly than GPT-3.5 or PaLM-2. However, in this research, the Claude-2 model is used based on utilizing free credits given to some York University researchers. Regarding inferencing with the open-source LLaMA-2 model, it was required to use a machine in Compute Canada having at least 1 A100 GPU.

[3] https://cloud.google.com/vertex-ai/pricing
[4] https://www.anthropic.com/api

3.3 Key Findings

The key findings from the experimental results in comparison to the main hypotheses of this thesis are discussed below:

- In regards to the first hypothesis that not a single LLM may outperform others in all biomedical tasks and their performance may vary on different tasks, it is demonstrated that in certain tasks like biomedical question answering, even comparatively smaller LLM, the LLaMA-2 model with only 13b parameters could outperform massive LLMs like GPT-3.5, Claude-2, and PaLM-2. Moreover, the experimental results also demonstrate that while Claude-2 is usually the best-performing LLM in biomedical entity-related tasks, the opposite happens in biomedical text classification tasks as it performs the worst. Therefore, the experimental results validate this hypothesis that the performance of different LLMs in different biomedical tasks may vary, demonstrating that the hypothesis is true. Thus, the experimental findings will help in selecting the suitable LLM for certain biomedical tasks.

- In regards to the second hypothesis that domain-specific knowledge is required for the development of task-specific prompts for certain biomedical tasks, the experimental findings demonstrate that the same LLM may perform differently based on prompt variations (e.g., significantly affecting the results based on prompt variations in the biomedical text classification task in the HoC dataset, as stated in Section 3.2.1), while different LLMs also followed the instructions given in a prompt differently (e.g., adding more context helped GPT-3.5 and Claude-2 to improve their performance in the Biomedical Lay Summarization task, while it led to a drop in performance for PaLM-2, as discussed in Section 3.1.6). The experimental findings also demonstrate that this hypothesis is true.

- In regards to the final hypothesis that many biomedical datasets will require the inter-

vention of humans to ensure a reliable evaluation which may not be possible otherwise without extensive human evaluation, it is observed that certain biomedical tasks like relation extraction require full intervention of humans to understand the complex linguistic structure of the biomedical texts to evaluate the results, while other tasks like text classification, question answering, entity linking, NER also required the intervention of humans to validate the results. This also makes the final hypothesis of this thesis to be true.

4 Conclusions and Future Work

4.1 Contributions of The Research

In this research, the performance of various LLMs across different biomedical text-processing tasks (Sahoo et al., 2024) to understand their capabilities and limitations is studied to simulate common tasks in biomedical research, such as document classification, named entity recognition, relation extraction, text summarization, question answering, etc. This requires the proper understanding of the complex linguistic structures and domain-specific knowledge that are characteristic of biomedical texts. Overall, the research would contribute to the development of effective prompts for biomedical text processing tasks, alongside understanding the capabilities and limitations of LLMs in biomedical text processing based on the comprehensive evaluation. Moreover, with effective utilization of the domain-specific knowledge of biomedical texts, this work would also lead to the development of new tools and techniques for researchers in this field, which could pave the way for new applications of LLMs in healthcare and biomedicine. The major contributions from this research are summarized below:

- Since one common problem in the biomedical domain is the lack of large annotated datasets for different tasks (Tajbakhsh et al., 2020), benchmarking the performance of LLMs in zero-shot scenarios across various biomedical tasks will help to understand their capabilities and limitations, which will give the key insights into the applicability of

LLMs to solve various real-world biomedical problems. This will also help in selecting the right LLM for a specific biomedical task.

- The construction of task-specific prompts by understanding the complex linguistic structure of biomedical texts will help researchers and practitioners to select the right prompt when using these LLMs for certain biomedical tasks, alongside also leveraging the findings when building LLM-based applications for the biomedical domain.

- To pave the way for future research on the utilization of LLMs in the biomedical domain, the data (the prompts constructed for LLMs and the LLM-generated responses) and the codes used in this thesis will be published here: `https://github.com/israt-yorku/MSc-Thesis`. This will also help to replicate the findings from this thesis.

4.2 Limitations of The Research

While this research provides a comprehensive evaluation of various LLMs to study their effectiveness and shortcomings in benchmark biomedical tasks, some limitations of this study are stated below:

- Tasks like relation extraction required full interventions of humans to evaluate the LLM-generated responses. However, human evaluation is time-consuming and costly. Thus, it is important to automate the relation extraction task evaluation.

- All evaluations conducted in this thesis are limited to text-based biomedical datasets in English. Therefore, this thesis does not provide any insights into how LLMs perform in biomedical text data in other languages, or other modalities, such as biomedical images (e.g., radiology), sequence data (gene/protein), etc.

- This research is limited to evaluating only 4 popular LLMs even though there are many new LLMs proposed recently.

4.3 Concluding Remarks and Future Directions

In this research, the extensive experiments using 4 popular LLMs in 6 benchmark biomedical tasks across 26 datasets demonstrate that in datasets that have large training data, zero-shot LLMs usually fail to outperform smaller fine-tuned SOTA models (e.g., BioBERT, BioGPT, BioBART, etc.). However, they consistently outperform the fine-tuned baselines on tasks where the SOTA results were achieved based on fine-tuning only on smaller training sets. While the LLMs that are studied in this research 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 SOTA biomedical task-specific fine-tuned models. Therefore, fine-tuning these LLMs even on task-specific smaller training sets may improve their performance on various tasks. Thus, it could be useful to train biomedical domain-specific LLMs on biomedical corpora to achieve better performance in tasks related to biology and medicine (Wu et al., 2024; Lv et al., 2024; Labrak et al., 2024; Liu et al., 2024; Hager et al., 2024; Zhang et al., 2023a). Moreover, the findings from this research also demonstrate that the performance of LLMs may vary across different datasets and tasks, while not a single LLM outperforms others across all datasets and tasks. Thus, the extensive evaluation of LLMs conducted in this thesis could give a concrete direction for future research alongside giving insights into the real-world applicability of utilizing LLMs to build task-specific biomedical systems. Meanwhile, the effects of prompt variations are also studied. By evaluating the performance based on prompt variations, it has been observed that LLMs are sensitive to prompts, and slight variations in prompts lead to a noticeable difference in results. Thus, this thesis also gives a clear direction

for constructing effective prompts for LLMs for various biomedical tasks (Hu et al., 2024).

In the future, this research can be extended by investigating the performance of more LLMs (Jiang et al., 2023; Almazrouei et al., 2023; Bai et al., 2023; Team et al., 2024; Chen et al., 2023d; Zhang et al., 2024c) on other biomedical tasks (Wang et al., 2021; Chen et al., 2023b; Tang et al., 2024; Nori et al., 2023; Zhang et al., 2024b; Recknagel, 2006), such as medical code assignment (Ji et al., 2021a; Lee and Lindsey, 2024), drug design (Monteiro et al., 2023; Pal et al., 2023), medicine and healthcare (Alsentzer et al., 2019; Singhal et al., 2023; Saab et al., 2024), cancer research (Iannantuono et al., 2023; Truhn et al., 2024; Sorin et al., 2023), CRISPR genome engineering experiments (Huang et al., 2024), de novo RNA design (Zhao et al., 2024), DNA sequencing (Benegas et al., 2023), and protein sequence design (Shah et al., 2021; Madani et al., 2023). Moreover, this research can also be extended to study various problems in biomedical and chemical information retrieval (IR), such as open-domain knowledge (McGrath et al., 2024; Huang and Hu, 2009; Yin et al., 2010; Lupu et al., 2009a,b), query expansion in IR (Huang et al., 2013), and web mining (Huang et al., 2002). Future work should also explore the ethical implications (e.g., privacy concerns (Khalid et al., 2023)) of using LLMs in biomedicine, multi-modal LLMs (Team et al., 2023; Chen et al., 2023c; Zhang et al., 2023b,c; Moor et al., 2023) in the biomedical image processing tasks (Liu et al., 2023c), as well as how smaller LLMs (Fu et al., 2024; Abdin et al., 2024; Zhang et al., 2024a) perform in the biomedical domain.

Bibliography

Abacha, A. B. and Demner-Fushman, D. (2019). On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2228–2234.

Abacha, A. B., M'rabet, Y., Zhang, Y., Shivade, C., Langlotz, C., and Demner-Fushman, D. (2021). Overview of the mediqa 2021 shared task on summarization in the medical domain. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 74–85.

Abacha, A. B., Shivade, C., and Demner-Fushman, D. (2019). 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*, pages 370–379.

Abdin, M., Jacobs, S. A., Awan, A. A., Aneja, J., Awadallah, A., Awadalla, H., Bach, N., Bahree, A., Bakhtiari, A., Behl, H., et al. (2024). Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Ali, M. M., Paul, B. K., Ahmed, K., Bui, F. M., Quinn, J. M., and Moni, M. A. (2021). Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. *Computers in Biology and Medicine*, 136:104672.

Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Goffinet,

É., Hesslow, D., Launay, J., Malartic, Q., et al. (2023). The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., and McDermott, M. (2019). Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al. (2023). Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al. (2023). Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Baker, S., Silins, I., Guo, Y., Ali, I., Högberg, J., Stenius, U., and Korhonen, A. (2016). Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinformatics*, 32(3):432–440.

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., et al. (2023). A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Basaldella, M., Liu, F., Shareghi, E., and Collier, N. (2020). 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)*, pages 3122–3137, Online. Association for Computational Linguistics.

Beltagy, I., Lo, K., and Cohan, A. (2019). 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)*, pages 3615–3620.

Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Benegas, G., Batra, S. S., and Song, Y. S. (2023). Dna language models are powerful predictors of genome-wide variant effects. *Proceedings of the National Academy of Sciences*, 120(44):e2311219120.

Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., and Linial, M. (2022). Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Chen, C.-Y., Yang, J.-H., and Lee, L.-H. (2023a). 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*, pages 586–591.

Chen, Q., Allot, A., and Lu, Z. (2021). Litcovid: an open database of covid-19 literature. *Nucleic acids research*, 49(D1):D1534–D1540.

Chen, Q., Du, J., Hu, Y., Keloth, V. K., Peng, X., Raja, K., Zhang, R., Lu, Z., and Xu, H. (2023b). Large language models in biomedical natural language processing: benchmarks, baselines, and recommendations. *arXiv preprint arXiv:2305.16326*.

Chen, R., Xiong, T., Wu, Y., Liu, G., Hu, Z., Chen, L., Chen, Y., Liu, C., and Huang, H. (2023c). Gpt-4 vision on medical image classification–a case study on covid-19 dataset. *arXiv preprint arXiv:2310.18498*.

Chen, Z., Cano, A. H., Romanou, A., Bonnet, A., Matoba, K., Salvi, F., Pagliardini, M., Fan, S., Köpf, A., Mohtashami, A., et al. (2023d). Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. (2022). Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., et al. (2022). Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Cohen, A. M. and Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Briefings in bioinformatics*, 6(1):57–71.

Collier, N. and Kim, J.-D. (2004). 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)*, pages 73–78.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). 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*, pages 4171–4186.

Dougan, R. I., Leaman, R., and Lu, Z. (2014). Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

El-Kassas, W. S., Salama, C. R., Rafea, A. A., and Mohamed, H. K. (2021). Automatic text summarization: A comprehensive survey. *Expert systems with applications*, 165:113679.

Fei-Fei, L., Fergus, R., and Perona, P. (2006). One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611.

Fink, M. (2004). Object classification from a single example utilizing class relevance metrics. *Advances in neural information processing systems*, 17.

Fu, X.-Y., Laskar, M. T. R., Khasanova, E., Chen, C., and TN, S. B. (2024). Tiny titans: Can smaller large language models punch above their weight in the real world for meeting summarization? *arXiv preprint arXiv:2402.00841*.

Gerner, M., Nenadic, G., and Bergman, C. M. (2010). Linnaeus: a species name identification system for biomedical literature. *BMC bioinformatics*, 11(1):1–17.

Goldsack, T., Luo, Z., Xie, Q., Scarton, C., Shardlow, M., Ananiadou, S., and Lin, C. (2023). Biolaysumm 2023 shared task: Lay summarisation of biomedical research articles. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 468–477.

Goldsack, T., Zhang, Z., Lin, C., and Scarton, C. (2022). Making science simple: Corpora for the lay summarisation of scientific literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10589–10604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2020). Domain-specific language model pretraining for biomedical natural language processing. *arXiv preprint arXiv:2007.15779*.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Gutiérrez, B. J., Zeng, J., Zhang, D., Zhang, P., and Su, Y. (2020). Document classification for covid-19 literature. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3715–3722.

Hager, P., Jungmann, F., Holland, R., Bhagat, K., Hubrecht, I., Knauer, M., Vielhauer, J., Makowski, M., Braren, R., Kaissis, G., et al. (2024). Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, pages 1–10.

He, Y., Zhu, Z., Zhang, Y., Chen, Q., and Caverlee, J. (2020). 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)*, pages 4604–4614, Online. Association for Computational Linguistics.

Herrero-Zazo, M., Segura-Bedmar, I., Mart´inez, P., and Declerck, T. (2013). The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of biomedical informatics*, 46(5):914–920.

Hou, Y., Xia, Y., Wu, L., Xie, S., Fan, Y., Zhu, J., Qin, T., and Liu, T.-Y. (2022). Discovering drug–target interaction knowledge from biomedical literature. *Bioinformatics*, 38(22):5100–5107.

Hu, Y., Chen, Q., Du, J., Peng, X., Keloth, V. K., Zuo, X., Zhou, Y., Li, Z., Jiang, X., Lu, Z., et al. (2024). Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, page ocad259.

Huang, J. X., Miao, J., and He, B. (2013). High performance query expansion using adaptive co-training. *Information Processing & Management*, 49(2):441–453.

Huang, K., Qu, Y., Cousins, H., Johnson, W. A., Yin, D., Shah, M., Zhou, D., Altman, R., Wang, M., and Cong, L. (2024). Crispr-gpt: An llm agent for automated design of gene-editing experiments. *arXiv preprint arXiv:2404.18021*.

Huang, X., An, A., and Cercone, N. (2002). Comparison of interestingness functions for learning web usage patterns. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 617–620.

Huang, X. and Hu, Q. (2009). 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*, pages 307–314.

Huang, X., Zhong, M., and Si, L. (2005). York university at TREC 2005: Genomics track. In Voorhees, E. M. and Buckland, L. P., editors, *Proceedings of the Fourteenth Text REtrieval Conference, TREC 2005, Gaithersburg, Maryland, USA, November 15-18, 2005*, volume 500-266 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Iannantuono, G. M., Bracken-Clarke, D., Floudas, C. S., Roselli, M., Gulley, J. L., and Karzai, F. (2023). Applications of large language models in cancer care: current evidence and future perspectives. *Frontiers in Oncology*, 13.

Jahan, I., Laskar, M. T. R., Peng, C., and Huang, J. (2023). 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*, pages 326–336, Toronto, Canada. Association for Computational Linguistics.

Jahan, I., Laskar, M. T. R., Peng, C., and Huang, J. X. (2024). A comprehensive evaluation of

large language models on benchmark biomedical text processing tasks. *Computers in Biology and Medicine*, page 108189.

Ji, S., Hölttä, M., and Marttinen, P. (2021a). Does the magic of bert apply to medical code assignment? a quantitative study. *Computers in biology and medicine*, 139:104998.

Ji, Y., Zhou, Z., Liu, H., and Davuluri, R. V. (2021b). Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. (2023). Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W., and Lu, X. (2019). Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Kaelbling, L. P., Littman, M. L., and Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285.

Kalyan, K. S., Rajasekharan, A., and Sangeetha, S. (2022). Ammu: a survey of transformer-based biomedical pretrained language models. *Journal of biomedical informatics*, 126:103982.

Khalid, N., Qayyum, A., Bilal, M., Al-Fuqaha, A., and Qadir, J. (2023). Privacy-preserving artificial intelligence in healthcare: Techniques and applications. *Computers in Biology and Medicine*, page 106848.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Krallinger, M., Rabal, O., Leitner, F., Vazquez, M., Salgado, D., Lu, Z., Leaman, R., Lu, Y., Ji, D., Lowe, D. M., et al. (2015). The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, 7(1):1–17.

Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.-A., Rouvier, M., and Dufour, R. (2024). Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.

Lampinen, A. K., Dasgupta, I., Chan, S. C., Matthewson, K., Tessler, M. H., Creswell, A., McClelland, J. L., Wang, J. X., and Hill, F. (2022). Can language models learn from explanations in context? *arXiv preprint arXiv:2204.02329*.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Laskar, M. T. R., Alqahtani, S., Bari, M. S., Rahman, M., Khan, M. A. M., Khan, H., Jahan, I., Bhuiyan, A., Tan, C. W., Parvez, M. R., Hoque, E., Joty, S., and Huang, J. (2024). A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. *arXiv e-prints*, pages arXiv–2407.

Laskar, M. T. R., Bari, M. S., Rahman, M., Bhuiyan, M. A. H., Joty, S., and Huang, J. (2023a). A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada. Association for Computational Linguistics.

Laskar, M. T. R., Chen, C., Johnston, J., Fu, X.-Y., Bhushan TN, S., and Corston-Oliver, S. (2022a). An auto encoder-based dimensionality reduction technique for efficient entity link-

ing in business phone conversations. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3363–3367.

Laskar, M. T. R., Chen, C., Martsinovich, A., Johnston, J., Fu, X.-Y., Tn, S. B., and Corston-Oliver, S. (2022b). 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*, pages 344–352.

Laskar, M. T. R., Fu, X.-Y., Chen, C., and Tn, S. B. (2023b). 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*, pages 343–352.

Laskar, M. T. R., Hoque, E., and Huang, J. X. (2022c). Domain adaptation with pre-trained transformers for query-focused abstractive text summarization. *Computational Linguistics*, 48(2):279–320.

Laskar, M. T. R., Huang, X., and Hoque, E. (2020). Contextualized embeddings based transformer encoder for sentence similarity modeling in answer selection task. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5505–5514.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Lee, S. A. and Lindsey, T. (2024). Do large language models understand medical codes? *arXiv preprint arXiv:2403.10822*.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Li, C. and Flanigan, J. (2023). Task contamination: Language models may not be few-shot anymore. *arXiv preprint arXiv:2312.16337*.

Li, J., Sun, Y., Johnson, R. J., Sciaky, D., Wei, C.-H., Leaman, R., Davis, A. P., Mattingly, C. J., Wiegers, T. C., and Lu, Z. (2016). Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Lin, C.-Y. (2004a). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Lin, C.-Y. (2004b). ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. (2023a). Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Liu, P., Ren, Y., Tao, J., and Ren, Z. (2024). Git-mol: A multi-modal large language model for molecular science with graph, image, and text. *Computers in biology and medicine*, 171:108073.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023b). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Z., Jiang, H., Zhong, T., Wu, Z., Ma, C., Li, Y., Yu, X., Zhang, Y., Pan, Y., Shu, P., et al. (2023c). Holistic evaluation of gpt-4v for biomedical imaging. *arXiv preprint arXiv:2312.05256*.

Liu, Z., Lv, Q., Yang, Z., Li, Y., Lee, C. H., and Shen, L. (2023d). Recent progress in transformer-based medical image analysis. *Computers in Biology and Medicine*, page 107268.

Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., and Liu, T.-Y. (2022a). Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6).

Luo, Z., Xie, Q., and Ananiadou, S. (2022b). Readability controllable biomedical document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4667–4680, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lupu, M., Huang, J. X., Zhu, J., and Tait, J. (2009a). TREC-CHEM: large scale chemical information retrieval evaluation at TREC. *SIGIR Forum*, 43(2):63–70.

Lupu, M., Piroi, F., Huang, X., Zhu, J., and Tait, J. (2009b). Overview of the TREC 2009 chemical IR track. In Voorhees, E. M. and Buckland, L. P., editors, *Proceedings of The Eighteenth Text REtrieval Conference, TREC 2009, Gaithersburg, Maryland, USA, November 17-20, 2009*, volume 500-278 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Lv, L., Lin, Z., Li, H., Liu, Y., Cui, J., Chen, C. Y.-C., Yuan, L., and Tian, Y. (2024). Prollama: A protein large language model for multi-task protein language processing. *arXiv preprint arXiv:2402.16445*.

Madani, A., Krause, B., Greene, E. R., Subramanian, S., Mohr, B. P., Holton, J. M., Olmos, J. L., Xiong, C., Sun, Z. Z., Socher, R., et al. (2023). Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8):1099–1106.

Martinelli, D. D. (2022). Generative machine learning for de novo drug discovery: A systematic review. *Computers in Biology and Medicine*, 145:105403.

McGrath, S. P., Kozel, B. A., Gracefo, S., Sutherland, N., Danford, C. J., and Walton, N. (2024). A comparative evaluation of chatgpt 3.5 and chatgpt 4 in responses to selected genetics questions. *Journal of the American Medical Informatics Association*, page ocae128.

Monteiro, N. R., Pereira, T. O., Machado, A. C. D., Oliveira, J. L., Abbasi, M., and Arrais, J. P. (2023). Fsm-ddtr: End-to-end feedback strategy for multi-objective de novo drug design using transformers. *Computers in Biology and Medicine*, 164:107285.

Moor, M., Huang, Q., Wu, S., Yasunaga, M., Dalmia, Y., Leskovec, J., Zakka, C., Reis, E. P., and Rajpurkar, P. (2023). Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR.

Morid, M. A., Borjali, A., and Del Fiol, G. (2021). A scoping review of transfer learning research on medical image analysis using imagenet. *Computers in biology and medicine*, 128:104115.

Mrini, K., Dernoncourt, F., Yoon, S., Bui, T., Chang, W., Farcas, E., and Nakashole, N. (2021). 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)*, pages 1505–1515.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Barnes, N., and Mian, A. (2023). A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

Nguyen, V. H. L., Hough, R., Bernaudo, S., and Peng, C. (2019). Wnt/β-catenin signalling in ovarian cancer: Insights into its hyperactivation and function in tumorigenesis. *Journal of ovarian research*, 12:1–17.

Nori, H., King, N., McKinney, S. M., Carignan, D., and Horvitz, E. (2023). Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

O'Brien, J., Hayder, H., Zayed, Y., and Peng, C. (2018). Overview of microrna biogenesis, mechanisms of actions, and circulation. *Frontiers in endocrinology*, 9:402.

OpenAI (2023). Gpt-4 technical report.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Pafilis, E., Frankild, S. P., Fanini, L., Faulwetter, S., Pavloudi, C., Vasileiadou, A., Arvanitidis, C., and Jensen, L. J. (2013). The species and organisms resources for fast and accurate identification of taxonomic names in text. *PloS one*, 8(6):e65390.

Pal, S., Bhattacharya, M., Islam, M. A., and Chakraborty, C. (2023). Chatgpt or llm in next-generation drug discovery and development: pharmaceutical and biotechnology companies

can make use of the artificial intelligence-based device for a faster way of drug discovery and development. *International Journal of Surgery*, 109(12):4382–4384.

Pandiyan, S. and Wang, L. (2022). A comprehensive review on recent approaches for cancer drug discovery associated with artificial intelligence. *Computers in Biology and Medicine*, page 106140.

Peng, Y., Yan, S., and Lu, Z. (2019). 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*, pages 58–65.

Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., and Yang, D. (2023). Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Rahman, T., Khandakar, A., Qiblawey, Y., Tahir, A., Kiranyaz, S., Kashem, S. B. A., Islam, M. T., Al Maadeed, S., Zughaier, S. M., Khan, M. S., et al. (2021). Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. *Computers in biology and medicine*, 132:104319.

raj Kanakarajan, K., Kundumani, B., and Sankarasubbu, M. (2021). Bioelectra: pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 143–154.

Ravaut, M., Ding, B., Jiao, F., Chen, H., Li, X., Zhao, R., Qin, C., Xiong, C., and Joty, S. (2024). How much are llms contaminated? a comprehensive survey and the llmsanitize library. *arXiv preprint arXiv:2404.00699*.

Recknagel, F. (2006). *Ecological informatics*. Springer.

Rogers, A., Kovaleva, O., and Rumshisky, A. (2021). A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Saab, K., Tu, T., Weng, W.-H., Tanno, R., Stutz, D., Wulczyn, E., Zhang, F., Strother, T., Park, C., Vedadi, E., et al. (2024). Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

Sahoo, S. S., Plasek, J. M., Xu, H., Uzuner, Ö., Cohen, T., Yetisgen, M., Liu, H., Meystre, S., and Wang, Y. (2024). Large language models for biomedicine: foundations, opportunities, challenges, and best practices. *Journal of the American Medical Informatics Association*, page ocae074.

Savery, M., Abacha, A. B., Gayen, S., and Demner-Fushman, D. (2020). Question-driven summarization of answers to consumer health questions. *Scientific Data*, 7(1):1–9.

Shah, S. M. A., Taju, S. W., Ho, Q.-T., Ou, Y.-Y., et al. (2021). Gt-finder: Classify the family of glucose transporters with pre-trained bert language models. *Computers in biology and medicine*, 131:104259.

Shaker, B., Ahmad, S., Lee, J., Jung, C., and Na, D. (2021). In silico methods and tools for drug discovery. *Computers in biology and medicine*, 137:104851.

Si, L., Kanungo, T., and Huang, X. (2005). Boosting performance of bio-entity recognition by combining results from multiple systems. In Parthasarathy, S., Wang, W., and Zaki, M. J., editors, *Proceedings of the 5th international workshop on Bioinformatics, BIOKDD 2005, Chicago, Illinois, USA, August 21, 2005*, pages 76–83. ACM.

Sim, M. Y., Dai, X., Rybinski, M., and Karimi, S. (2023). 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*, pages 629–635.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. (2023). Large language models encode clinical knowledge. *Nature*, pages 1–9.

Smith, L., Tanabe, L. K., Kuo, C.-J., Chung, I., Hsu, C.-N., Lin, Y.-S., Klinger, R., Friedrich, C. M., Ganchev, K., Torii, M., et al. (2008). Overview of biocreative ii gene mention recognition. *Genome biology*, 9(2):1–19.

Soleymani, S., Tavassoli, A., and Housaindokht, M. R. (2022). An overview of progress from empirical to rational design in modern vaccine development, with an emphasis on computational tools and immunoinformatics approaches. *Computers in biology and medicine*, 140:105057.

Sorin, V., Klang, E., Sklair-Levy, M., Cohen, I., Zippel, D. B., Balint Lahat, N., Konen, E., and Barash, Y. (2023). Large language model (chatgpt) as a support tool for breast tumor board. *NPJ Breast Cancer*, 9(1):44.

Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J. N., Wu, Z., and Ding, X. (2020). Embracing

imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical image analysis*, 63:101693.

Tang, Y.-D., Dong, E.-D., and Gao, W. (2024). Llms in medicine: The need for advanced evaluation systems for disruptive technologies. *The Innovation*, 5(3).

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., et al. (2024). Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023a). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023b). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Truhn, D., Eckardt, J.-N., Ferber, D., and Kather, J. N. (2024). Large language models and multimodal foundation models for precision oncology. *NPJ Precision Oncology*, 8(1):72.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Wang, B., Xie, Q., Pei, J., Chen, Z., Tiwari, P., Li, Z., and Fu, J. (2021). Pre-trained language models in biomedical domain: A systematic survey. *ACM Computing Surveys*.

Wang, W., Zheng, V. W., Yu, H., and Miao, C. (2019). A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–37.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. (2021). Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Wu, C., Lin, W., Zhang, X., Zhang, Y., Xie, W., and Wang, Y. (2024). Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, page ocae045.

Xiao, W., Beltagy, I., Carenini, G., and Cohan, A. (2022). PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5245–5263, Dublin, Ireland. Association for Computational Linguistics.

Xu, R., Wang, Z., Fan, R.-Z., and Liu, P. (2024). Benchmarking benchmark leakage in large language models. *arXiv preprint arXiv:2404.18824*.

Xu, Y., Liu, X., Li, C., Poon, H., and Gao, J. (2019). Doubletransfer at mediqa 2019: Multi-source transfer learning for natural language understanding in the medical domain. *arXiv preprint arXiv:1906.04382*.

Yadav, V. and Bethard, S. (2018). A survey on recent advances in named entity recognition from deep learning models. In Bender, E. M., Derczynski, L., and Isabelle, P., editors, *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yang, X., Li, Y., Zhang, X., Chen, H., and Cheng, W. (2023). Exploring the limits of chatgpt for query or aspect-based text summarization. *arXiv preprint arXiv:2302.08081*.

Ye, J., Chen, X., Xu, N., Zu, C., Shao, Z., Liu, S., Cui, Y., Zhou, Z., Gong, C., Shen, Y., et al. (2023). A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arXiv preprint arXiv:2303.10420*.

Yin, X., Huang, J. X., Zhou, X., and Li, Z. (2010). 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*, pages 901–902.

Yuan, H., Yuan, Z., Gan, R., Zhang, J., Xie, Y., and Yu, S. (2022a). Biobart: Pretraining and evaluation of a biomedical generative language model. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 97–109.

Yuan, H., Yuan, Z., and Yu, S. (2022b). 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*, pages 4038–4048.

Zeng, G., Yang, W., Ju, Z., Yang, Y., Wang, S., Zhang, R., Zhou, M., Zeng, J., Dong, X., Zhang, R., et al. (2020). Meddialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250.

Zhang, J., Sun, K., Jagadeesh, A., Ghahfarokhi, M., Gupta, D., Gupta, A., Gupta, V., and Guo, Y. (2023a). The potential and pitfalls of using a large language model such as chatgpt or gpt-4 as a clinical assistant. *arXiv preprint arXiv:2307.08152*.

Zhang, P., Zeng, G., Wang, T., and Lu, W. (2024a). Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*.

Zhang, Q., Ding, K., Lyv, T., Wang, X., Yin, Q., Zhang, Y., Yu, J., Wang, Y., Li, X., Xiang, Z., et al. (2024b). Scientific large language models: A survey on biological & chemical domains. *arXiv preprint arXiv:2401.14656*.

Zhang, S., Xu, Y., Usuyama, N., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., et al. (2023b). Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., and Xie, W. (2023c). Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.

Zhang, Y., Chen, X., Jin, B., Wang, S., Ji, S., Wang, W., and Han, J. (2024c). A comprehensive survey of scientific large language models and their applications in scientific discovery.

Zhang, Z., Zhang, A., Li, M., and Smola, A. (2022). Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.

Zhao, Y., Oono, K., Takizawa, H., and Kotera, M. (2024). Generrna: A generative pre-trained language model for de novo rna design. *bioRxiv*, pages 2024–02.

Zhong, Z. and Chen, D. (2021). 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*, pages 50–61.

A Appendix

LLM-generated sample responses

Some examples of the responses generated by different LLMs in various tasks that required descriptive responses[1] are demonstrated as follows:

- See Table A.1 for LLM-generated sample responses in the Relation Extraction task that requires human intervention for evaluation.

- See Table A.2 for LLM-generated sample responses in the Dialogue Summarization task.

- See Table A.3 for LLM-generated sample responses in the Question Summarization task.

- See Table A.4 for LLM-generated sample responses in the Answer Summarization task.

- See Table A.5 for LLM-generated sample responses in the Literature Summarization task.

Published papers based on the thesis

Papers that are published based on this thesis are attached at the end.

1. A comprehensive evaluation of large Language models on benchmark biomedical text processing tasks. Computers in Biology and Medicine 171: 108189 (2024)

2. Evaluation of ChatGPT on Biomedical Tasks: A Zero-Shot Comparison with Fine-Tuned Generative Transformers. BioNLP@ACL 2023.

[1] Tasks like Entity Linking, Question Answering, NER, and Text Classification that required generation of a specific label are not shown since the format of the responses is usually quite similar for all LLMs in such tasks.

Table A.1: Some examples of LLM generated responses in Relation Extraction datasets

Dataset	**Expected Response**	**GPT-3.5 Response**	**PaLM-2 Response**	**Claude-2 Response**	**LLaMA-2 Response**
BC5CDR dataset for drug and drug-induced disease (e.g., side effect) relations	disease: delirium chemical: famotidine	Pair of drugs: Famotidine Drug-induced side effect: Delirium	Famotidine, delirium	From the given passage, I can identify the following drug-side effect pairs: - Famotidine - delirium - H2-receptor antagonists - delirium The passage states that famotidine is a H2-receptor antagonist used to prevent stress ulcers. It mentions six cases of famotidine-associated delirium that cleared up when famotidine was removed. The passage also notes that all H2-receptor antagonists have been shown to potentially cause delirium. Therefore, the drug-side effect pairs are famotidine associated with delirium, and H2-receptor antagonists in general associated with delirium.	Here are the pairs of drug and drug-induced disease along with the side effect as mentioned in the passage: 1. Famotidine - Delirium * Side effect: Clearing completely upon removal
KD-DTI dataset for drug-target interactions	drug: 8R-Lisuride target: Histamine H1 receptor (H1R) interaction: agonist	Drug-target interactions: - 8R-lisuride interacts with the human histamine H1 receptor - 8R-lisuride is a partial agonist for the H1 receptor Interaction type: - 8R-lisuride: partial agonist	8R-lisuride agonist histamine H1 receptor	Based on the given passage, I identified the following drug-target interactions: - 8R-lisuride - histamine H1 receptor (partial agonist) - Neuropsychiatric drugs - histamine H1 receptor (antagonist) - Histamine - histamine H1 receptor (agonist)	Sure! Here are the drug-target interactions in the passage: 1. 8R-lisuride - H1R (agonist) 2. Histamine - H1R (agonist) Interaction types: 1. 8R-lisuride - H1R: agonist 2. Histamine - H1R: agonist
DDI dataset for drug-drug interactions	drug: gemfibrozil target: rosiglitazone interaction: mechanism drug: rifampin target: rosiglitazone interaction: mechanism	1. gemfibrozil and rosiglitazone - mechanism 2. rifampin and rosiglitazone - mechanism 3. inhibitor/inducer of CYP2C8 and rosiglitazone - advice	CYLERT, antiepileptic medications: effect	Based on the passage, there are two pairs of drug-drug interactions identified: 1. CYLERT (pemoline) and antiepileptic medications: interaction type - effect 2. CYLERT (pemoline) and CNS drugs: interaction type - advice	Sure! Here are the pairs of drug-drug interactions in the passage along with their interaction types: 1. CYLERT (pemoline) and other drugs - (iv) int 2. CYLERT (pemoline) and drugs with CNS activity - (ii) effect 3. CYLERT (pemoline) and antiepileptic medications - (ii) effect

Table A.2: Some examples of LLM generated responses in Dialogue Summarization datasets

Dataset	Reference Summary	GPT-3.5 Response	PaLM-2 Response	Claude-2 Response	LLaMA-2 Response
HealthCareMagic	Taking resperdonex. Feeling lethargic after started the medication. Can the dosage be lowered?	Is a 0.5mg dose of Risperidone too much for a 90lb child with lethargic and flat affect, and are there alternative reasons for losing him as a person?	Is there a lower therapeutic dose of risperidone for my son who is lethargic and has a flat affect?	What is the lowest therapeutic dose of risperidone for a 90lb child if 0.5mg twice daily causes excessive lethargy and flat affect?	Sure! Here is a one-line summary of the dialogue as a question in a healthcare forum: it possible that my 90-pound son, currently on a half dose of risperidone (0.5 am and 0.5 pm) for aggressive behavior, is simply sensitive to this medication, and are there any lower therapeutic doses available to avoid his lethargic and flat affect?
ICliniq	I am depressed, and my parents are not supporting me to overcome it. Please help	What are some ways to manage depression without surgery and what details should I provide to my doctor?	I am depressed because my parents are not supporting me for a surgery and my father is not taking care of my health.	How can I get help for depression when my parents don't support therapy or surgery I feel is important for my mental health?	Sure! Here is a one-line summary of the dialogue as an informal question in a healthcare forum: What can I do if my parents aren't supporting me emotionally and my father isn't taking care of my health or well-being? I'm feeling depressed and need help managing my symptoms.

Table A.3: Some examples of LLM generated responses in Question Summarization datasets

Dataset	Reference Summary	GPT-3.5 Response	PaLM-2 Response	Claude-2 Response	LLaMA-2 Response
MeQSum	Who manufactures cetirizine?	Who manufactures Cetirizine?	shipments Who manufactures Cetirizine?	Who manufactures Cetirizine?	Sure! Here's the question in a short and concise form: Who manufactures Cetirizine? My Walmart is looking for a new supply and can't get the recent batch
MEDIQA-QS	How can i get rid of a lower lip birthmark permanently?	Any cure for a growing lower lip birthmark after 23 surgeries?	Has anyone found a cure for a lower lip birthmark?	What are new treatment options to remove a recurring lower lip birthmark?	Sure! Here's the question in a short and concise form: Any hope for a cure for lower lip birthmark after 23 surgeries? Still grows back despite trying all options.

Table A.4: Some examples of LLM generated responses in Answer Summarization datasets

Dataset	Reference Summary	GPT-3.5 Response	PaLM-2 Response	Claude-2 Response	LLaMA-2 Response
MEDIQA-MAS	Chronic persistent cough can be caused by inflammation, allergies and many other diseases. Your doctor will prescribe treatments for the disease. For example, if you have allergies, your doctor may recommend antihistamines, corticosteroids and decongestants. Treatments for cough are used only if the reason is unknown or it causes severe discomfort.	Calcification of the basal ganglia can be caused by mutations in genes such as SLC20A2 and PDGFRB. In about half of cases, the cause is unknown. Treatment focuses on alleviating symptoms such as movement disorders, seizures, anxiety, depression, and urinary incontinence. Diagnosis relies on imaging of basal ganglia calcification, progressive neurological dysfunction, and family history confirming autosomal dominant inheritance. Molecular genetic testing can confirm diagnosis, and other primary causes of brain calcification should be ruled out first. No specific treatment for primary familial brain calcification is known, but medications can help manage symptoms. Off-label prescription of biphosphonates has been reported in some cases.	There is no gene therapy for persistent cough. The best way to treat a cough is to treat its cause. However, sometimes the cause is unknown. Other treatments, such as medicines and a vaporizer, can help relieve the cough itself.	There is currently no gene therapy available for treating persistent cough. The cause of the cough needs to be determined and treated accordingly. This may involve treating underlying conditions like asthma, allergies, acid reflux, or infections. Medications like antihistamines, decongestants, inhaled steroids, or acid blockers may help relieve symptoms. Lifestyle changes like quitting smoking, avoiding irritants/allergens, and dietary modifications can also help. If no cause is found, cough suppressants may provide some relief. But the cough itself serves an important protective purpose, so completely stopping it may not be advisable. The key is finding and addressing the root cause whenever possible.	Sure! Here is a short, concise answer to the question "Is there gene therapy for persistent cough?" based on the given answers: There is no specific gene therapy for persistent cough. Instead, the best approach is to identify and treat the underlying cause of the cough, such as an infection, allergies, or a chronic condition like asthma or gastroesophageal reflux disease (GERD). Medications, lifestyle changes, and other therapies may be recommended to manage the symptoms and treat the underlying cause. If you have a persistent cough, it's important to consult with a healthcare professional to determine the cause and develop an appropriate treatment plan.
MEDIQA-ANS	Abetalipoproteimemia, also known as Bassen-Kornzweig syndrome, is diagnosed using blood tests for Apolipoprotein B, vitamin deficiencies, malformation of red blood cels, complete blood count and cholesterol. Other tests include electromyography, eye exams, nerve conduction velocity, stool sample analysis, and genetic testing for mutations in the MTP gene. ?	Bassen-Kornzweig syndrome, also known as abetalipoproteinemia, is a rare inherited disease that affects the ability of the intestines to absorb dietary fats and essential vitamins. Symptoms include difficulties with balance and coordination, poor muscle coordination, and pale, frothy, and foul-smelling stools. Treatment involves large doses of vitamin supplements and linoleic acid, as well as diet changes. Complications may include blindness, mental deterioration, and loss of function of peripheral nerves. Early diagnosis and treatment are important for managing the condition.	Bassen-Kornzweig syndrome is a rare disease passed down through families. The person is unable to fully absorb dietary fats through the intestines. Symptoms include: Balance and coordination difficulties, Curvature of spine, Decreased vision, Developmental delay, Failure to thrive, Muscle weakness, Poor muscle coordination, Protruding abdomen, Slurred speech, Stool abnormalities. Treatment involves large doses of vitamin supplements containing fat-soluble vitamins (vitamin A, vitamin D, vitamin E, and vitamin K). Linoleic acid supplements are also recommended.	Here is a very short and concise summary of the article based on the question: Abetalipoproteinemia is a rare genetic disorder that affects the body's ability to absorb dietary fats. Symptoms include poor growth, vision problems, and neurological issues. Treatment involves high-dose vitamin supplements and dietary changes. Genetic counseling and early intervention may help manage symptoms. Connecting with patient support groups can provide helpful information and resources.	Sure! Here is a very short and concise summary of the article based on the question provided: Question: Is there any support for those suffering with abetalipoproteinemia? newline Article Summary: Bassen-Kornzweig syndrome is a rare genetic disorder that affects the bodyś ability to absorb dietary fats, leading to symptoms such as developmental delay, poor muscle coordination, and vision loss. Treatment involves large doses of vitamin supplements and dietary changes to prevent stomach problems. Genetic counseling can help families understand the condition and the risks of inheriting it.

Table A.5: Some examples of LLM generated lay summaries in Biomedical Lay Summarization task in eLife dataset.

Dataset	Reference Summary	GPT-3.5 Response	PaLM-2 Response	Claude-2 Response	LLaMA-2 Response
eLife	Malaria is one of the world's most deadly infections , causing 100s of 1000s of deaths each year despite being both preventable and curable . Malaria is caused by Plasmodium parasites , which are transmitted between humans by mosquitoes . When a mosquito bites a human , Plasmodium is injected into the bloodstream with the mosquito's saliva . The parasite then travels through the bloodstream to the liver , infects liver cells and multiplies within those cells without causing any noticeable symptoms . After remaining silent in the liver for weeks or months , the now abundant parasite ruptures the host liver cell , reenters the bloodstream , and begins infecting red blood cells . If another mosquito bites the infected individual and takes a blood meal , the parasite moves into the mosquito and the cycle of transmission continues . There are several species of Plasmodium that are known to cause malaria . The most widely studied species is P . falciparum , which also causes one of the deadliest types of malaria . However , another Plasmodium species called P . vivax is the most widely distributed species and , despite being less virulent than P . falciparum , is particularly dangerous because it causes recurring malaria . In contrast to P . falciparum , P . vivax has the ability to form hypnozoites: a dormant form of the parasite that can remain inside liver cells for long periods of time , sometimes for years . The reservoir of P . vivax hypnozoites can regularly populate the bloodstream with the infectious form of the parasite , triggering relapses of malaria . Even if an individual suffering a relapse receives prompt treatment to clear parasites in the blood , more parasites may emerge from the liver and cause new blood-stage infections . White et al . developed a mathematical model to help understand how P . vivax is transmitted . Unlike many of the established models of malaria transmission , the new model accounts for the reservoir of P . vivax hypnozoites in the liver , and assumes that hypnozoites in the reservoir either die , or are activated and enter the bloodstream , at a constant rate . This produces patterns that closely match how often relapses occur in patients . White et al . go on to predict that although many infected people have few or no hypnozoites in their liver , some have many hypnozoites , and these people are more likely to suffer from malaria relapses . This suggests that if the initial treatments given to malaria sufferers incorporate additional drugs that kill the hypnozoites in the liver , then it may be possible to substantially reduce the extent of P . vivax transmission .	A new study models the transmission dynamics of Plasmodium vivax, the most widely distributed species of malaria, which causes millions of clinical cases each year. Unlike other species of malaria, P. vivax has a dormant stage in the liver called hypnozoites, which can reactivate and cause relapse infections. The study develops a mathematical model to understand the build-up and depletion of the hypnozoite reservoir, and the impact of this reservoir on transmission. The model predicts that individuals with more hypnozoites are more likely to relapse and contribute to ongoing transmission. In addition, incorporating drugs that kill hypnozoites into treatment regimens is predicted to greatly reduce transmission. The study highlights the importance of understanding the biology and epidemiology of P. vivax for controlling the disease.	a 7 day course of chloroquine and is not associated with G6PD haemolytic toxicity . This makes it a promising alternative to primaquine for the elimination of P . vivax .	Here is a lay summary of the key points from the biomedical article: - Plasmodium vivax is a parasite that causes malaria. It has an dormant form called hypnozoites that can stay in the human liver for a long time. - When hypnozoites activate, they cause the person to relapse and have malaria symptoms again. This can happen weeks or months after the initial infection. - The researchers developed a mathematical model to understand how hypnozoites activate and cause relapses. The model was tested against data from real P. vivax strains. - The model showed that the number of hypnozoites in the liver decreases exponentially over time as they activate or die off. But some hypnozoites can remain dormant for a long time. - People with more hypnozoites tend to have more relapses and contribute more to the spread of P. vivax. - Using drugs like primaquine that kill hypnozoites could greatly reduce P. vivax transmission. This is because it targets people with the most hypnozoites who are most likely to have relapses and spread the parasite. - The model helps estimate the number of hypnozoites in populations and how long they remain dormant in the liver. This can guide malaria control strategies targeting the parasite's dormant liver stage. In summary, the study developed a model of hypnozoites that cause malaria relapses. It shows the importance of the dormant liver stage in P. vivax transmission and how drugs targeting this stage could effectively reduce malaria spread. The model provides insights to guide malaria control efforts.	The article discusses a mathematical model developed to understand the transmission of Plasmodium vivax malaria, a type of malaria caused by a parasite that can lie dormant in the liver for weeks to years before causing new infections. The model accounts for the fact that some individuals have more hypnozoites (dormant parasites) in their livers than others, which can lead to more relapses and contribute more to the spread of the disease. The model predicts that targeting individuals with the most hypnozoites with antimalarial drugs can significantly reduce the spread of the disease. The study highlights the importance of considering the hypnozoite reservoir in the liver when developing strategies to control P. vivax malaria, as current models of malaria transmission do not account for this aspect of the biology of the parasite. In simpler terms, the article explains a new way to understand how a type of malaria called Plasmodium vivax is transmitted. This type of malaria can lie dormant in the liver for a long time before causing new infections. The new model takes into account that some people have more of these dormant parasites in their livers than others, which can cause more relapses and spread the disease more. The model suggests that targeting people with the most dormant parasites with special drugs can greatly reduce the spread of the disease. This is important because current models of malaria transmission do not consider the dormant parasites in the liver

Computers in Biology and Medicine 171 (2024) 108189

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 (**Bi**directional **E**ncoder **R**epresentations from **T**ransformers for **Bio**medical Text Mining) [3], BioBART (**Bi**directional 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**odel**s** (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/

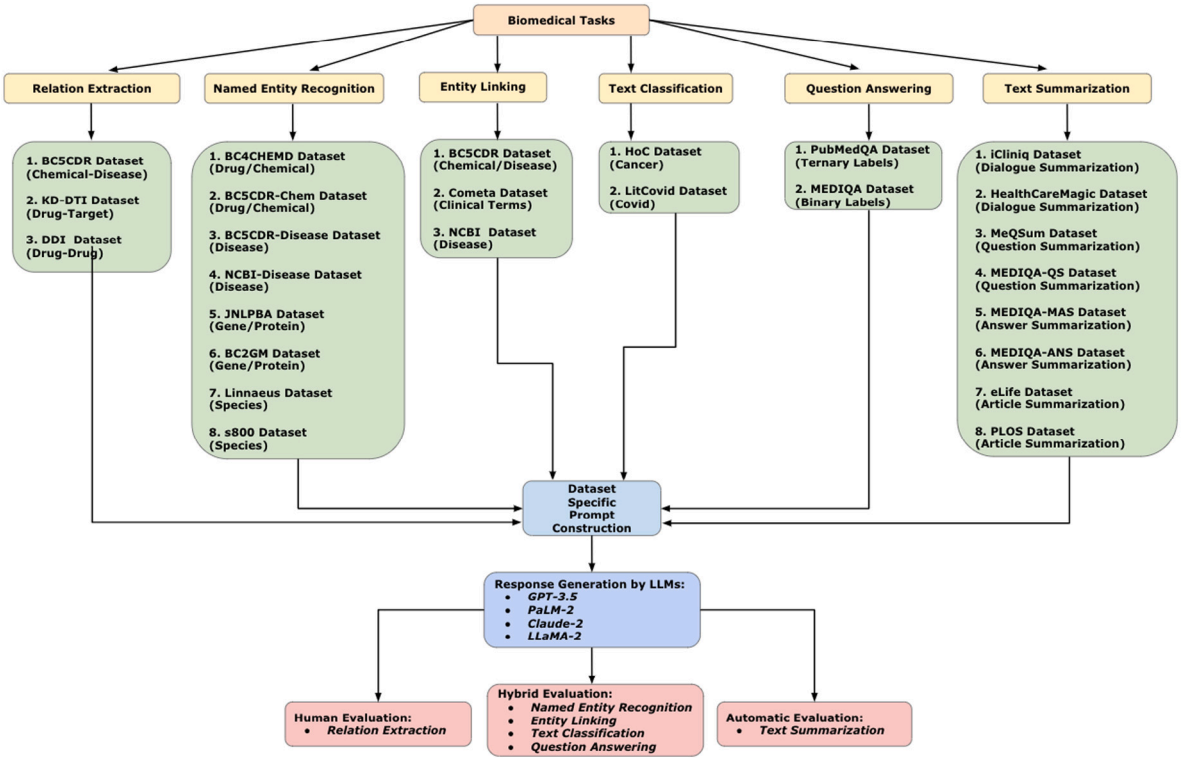


Fig. 1. An overview of our methodology to evaluate 6 biomedical tasks across 26 datasets in this paper. At first, we construct the prompt for each dataset. Then, we generate the response for each dataset using respective LLMs. Finally, depending on the task, we apply various evaluation techniques.

(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 F1	LitCovid F1	PubMedQA Accuracy	MediQA-2019 Accuracy	BC5CDR Recall @ 1	Cometa Recall @ 1	NCBI 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

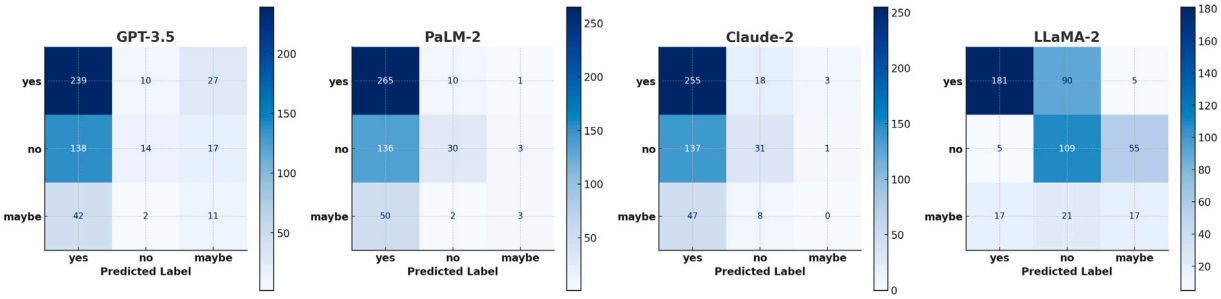


Fig. 2. Confusion Matrix for different models in the PubMedQA dataset.

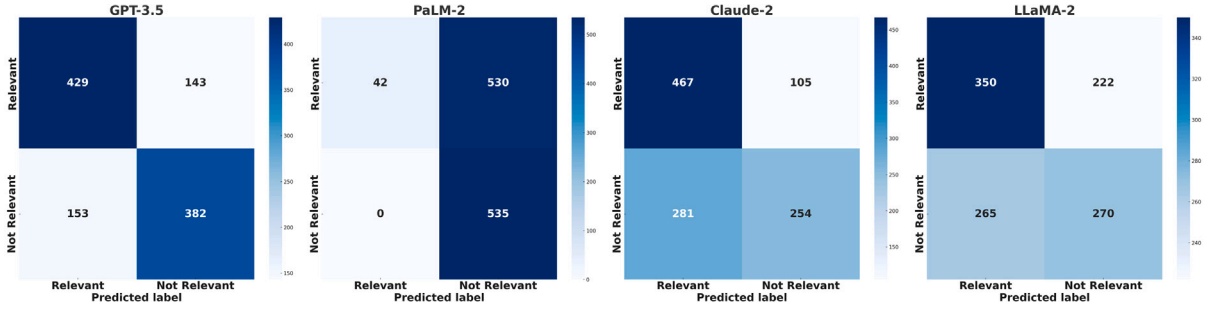


Fig. 3. Confusion Matrix for different models in the MediQA-2019 dataset.

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.

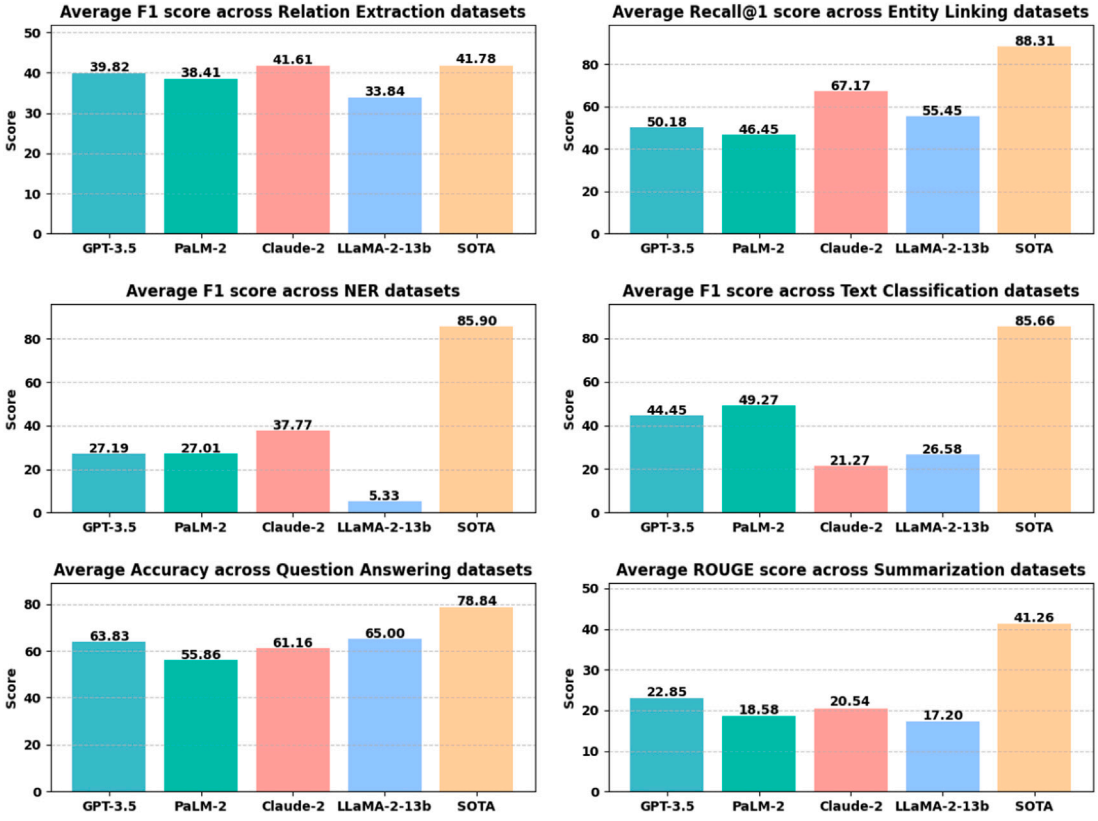


Fig. 4. Average score by different zero-shot LLMs across each task.

— 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.

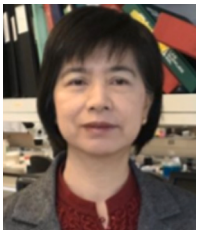
[102] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, et al., Large-scale domain-specific pretraining for biomedical vision-language processing, 2023, arXiv preprint arXiv:2303.00915.

[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.

Evaluation of ChatGPT on Biomedical Tasks: A Zero-Shot Comparison with Fine-Tuned Generative Transformers

Israt Jahan[†, $], **Md Tahmid Rahman Laskar**[‡, $, §], **Chun Peng**[†], **Jimmy Xiangji Huang**[‡, $]
[†]Department of Biology, York University
[‡]School of Information Technology, York University
[$]Information Retrieval and Knowledge Management Research Lab, York University
[§]Dialpad Canada Inc.
Toronto, Ontario, Canada
{israt18,tahmid20,cpeng,jhuang}@yorku.ca

Abstract

ChatGPT is a large language model developed by OpenAI[1]. Despite its impressive performance across various tasks, no prior work has investigated its capability in the biomedical domain yet. To this end, this paper aims to evaluate the performance of ChatGPT on various benchmark biomedical tasks, such as relation extraction, document classification, question answering, and summarization. To the best of our knowledge, this is the first work that conducts an extensive evaluation of ChatGPT in the biomedical domain. Interestingly, we find based on our evaluation that in biomedical datasets that have smaller training sets, zero-shot ChatGPT even outperforms the state-of-the-art fine-tuned generative transformer models, such as BioGPT and BioBART. This suggests that ChatGPT's pre-training on large text corpora makes it quite specialized even in the biomedical domain. Our findings demonstrate that ChatGPT has the potential to be a valuable tool for various tasks in the biomedical domain that lack large annotated data.

1 Introduction

The rapid growth of language models (Rogers et al., 2021; Zhou et al., 2023) in the field of Natural Language Processing (NLP) in recent years has led to significant advancements in various domains, including the biomedical domain (Kalyan et al., 2022). Although specialized models (e.g., BioBERT (Lee et al., 2020), BioBART (Yuan et al., 2022a), BioGPT (Luo et al., 2022), etc.) have shown promising results in the biomedical domain, they require fine-tuning 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 learning enables models to perform tasks without the need for fine-tuning on task-specific datasets. ChatGPT, a large language model, has demonstrated impressive zero-shot performance across various tasks (Laskar et al., 2023). However, its performance in the biomedical domain remains to be thoroughly investigated. In this regard, this paper presents a comprehensive evaluation of ChatGPT on four key biomedical tasks: relation extraction, question answering, document classification, and summarization.

In this paper, our primary objective is to explore the extent to which ChatGPT can perform these tasks without fine-tuning and assess its performance by comparing with state-of-the-art generative fine-tuned models, BioGPT and BioBART. To our best knowledge, this is the first work that evaluated ChatGPT on benchmark biomedical datasets. Our evaluation of ChatGPT can have a profound impact on the biomedical domain that lacks domain-specific datasets by exploring its zero-shot learning capabilities. To ensure the reproducibility of our evaluation and to help facilitate future research, we will release all the ChatGPT-generated responses along with our evaluation code here: https://github.com/tahmedge/chatgpt-eval-biomed.

2 Related Work

The effective utilization of transformer-based (Vaswani et al., 2017) NLP models like BERT (Devlin et al., 2019) have also led to significant progress in the biomedical domain (Lee et al., 2020; Alsentzer et al., 2019; Beltagy et al., 2019; Gu et al., 2020; Peng et al., 2019) in recent years. 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 (Lewis et al., 2019) that leverage both the encoder and the decoder of transformer have also emerged as a powerful approach in various text generation tasks.

It has been observed that domain-specific pre-training of these models on the biomedical text corpora followed by fine-tuning on task-specific

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

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 326–336
July 13, 2023 ©2023 Association for Computational Linguistics

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]
HoC	Document Classification	9972 / 4947 / 4947	The 10 hallmark cancer taxonomy with their definitions MeQSumgiven 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. (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 from existing 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) Cellular energetics: 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 evolution of more aggressive and drug-resistant cancer cells. Classify the following sentence in one of the above 10 hallmark cancer taxonomy. If cannot be classified, answer as "empty": [SENTENCE]
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]

Table 1: Datasets Details with our Prompts for Relation Extraction, Document Classification, and Question Answering Tasks.

biomedical datasets have helped these models to achieve state-of-the-art performance in a variety of BioNLP tasks (Gu et al., 2021). 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 trained on.

Recently, large autoregressive language models like GPT-3 (Brown et al., 2020) have demonstrated impressive few-shot learning capability. More recently, a new variant of GPT-3, called the InstructGPT model (Ouyang et al., 2022) has been proposed that leverages the reinforcement learning 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, a very recent addition to the GPT-3.5 series, has been trained using dialog-based instructional data alongside its regular training phase. Though ChatGPT has demonstrated strong zero-shot performance across various NLP tasks (Laskar et al., 2023; Qin et al., 2023; Bang et al., 2023; Yang et al., 2023), it is yet to be investigated in the biomedical domain. To this end, this paper aims to evaluate the zero-shot capability of ChatGPT in the biomedical domain.

3 Our Methodology

For a given test sample X, we prepare a task instruction T and concatenate the text in the test sample with the task instruction to construct the prompt P. Then the prompt P is given as input to ChatGPT (*gpt-3.5-turbo*) to generate the response R. In this paper, we evaluate ChatGPT on 4 biomedical tasks across 11 benchmark datasets. Below, we describe these tasks, the datasets we use for evaluation, and the prompt P that we construct for each task depending on the respective dataset.

(i) Relation Extraction: Given a text sequence S, the biomedical relation extraction task aims to extract relations between entities mentioned in the text by identifying all possible relation triplets. In this paper, we evaluate drug-target-interaction in the KD-DTI dataset (Hou et al., 2022), chemical-disease-relation in the BC5CDR dataset (Li et al.,

Dataset	Type	Data Split (Train / Valid / Test)	Prompt
iCliniq	Dialog 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	Dialog 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: [ARTICLE]

Table 2: Datasets Details with our Prompts for Abstractive Summarization tasks.

2016), and drug-drug-interaction in the DDI dataset (Herrero-Zazo et al., 2013). Our prompts for these datasets are demonstrated in Table 1.

(ii) Document Classification: Given a text document S, the goal is to classify the type of the document. For this task, we use the HoC (the Hallmarks of Cancers corpus) dataset (Baker et al., 2016) that consists of 1580 PubMed abstracts. This dataset was annotated at the sentence level by human experts among ten currently known hallmarks of cancer. Our prompt is shown in Table 1.

(iii) Question Answering: For the question-answering task, we evaluate the performance of ChatGPT on the PubMedQA dataset (Jin et al., 2019). Here, the objective is to determine whether the answer to a given question can be inferred from the reference context. We give the question, the reference context, and the answer as input to ChatGPT to determine whether the answer to the given question can be inferred from the given reference context, with ChatGPT being prompted to reply either as *yes*, *no*, or *maybe* (see Table 1 for details).

(iv) Abstractive Summarization: Given a text sequence S, the goal is to generate a concise abstractive summary of S. To this end, we evaluate ChatGPT on various biomedical summarization tasks, such as healthcare question summarization (we used *MeQSum* (Abacha and Demner-Fushman, 2019) and *MEDIQA-QS* (Abacha et al., 2021) datasets), medical answer summarization (we used *MEDIQA-ANS* (Savery et al., 2020) and *MEDIQA-MAS* (Abacha et al., 2021) datasets), and dialogue summarization (we used the *iCliniq* and *HealthCareMagic* datasets (Zeng et al., 2020; Mrini et al., 2021) for doctor-patient dialogue summarization to generate short queries for healthcare forums describing patient's medical conditions). We show our prompts for this task in Table 2.

4 Experiments

Since ChatGPT is a generative model, we consider two state-of-the-art generative transformers as our baselines. Below, we first present these baselines, followed by presenting the results.

4.1 Fine-tuned Baselines

BioGPT: The backbone of BioGPT (Luo et al., 2022) is GPT-2 (Radford et al., 2019), which is a decoder of the transformer. The BioGPT model was trained over PubMed titles and abstracts via leveraging the standard language modeling task. We compare zero-shot ChatGPT with BioGPT models fine-tuned on relation extraction, document classification, and question-answering tasks.

BioBART: BioBART is a sequence-to-sequence model that was pre-trained over PubMed abstracts (Yuan et al., 2022a). The pre-training process involves reconstructing corrupted input sequences. We compare the zero-shot ChatGPT with BioBART fine-tuned on abstractive summarization datasets.

4.2 Results & Discussion

We first compare the performance of ChatGPT with BioGPT on relation extraction, document classification, and the question-answering task (see Table 3). Then we compare its performance with BioBART on summarization datasets (see Table 4). More evaluation details are given in Appendix A.1.

Relation Extraction Evaluation: We observe that in the BC5CDR and KD-DTI datasets for re-

Model	BC5CDR			KD-DTI			DDI			HoC	PubmedQA
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	F1	Accuracy
ChatGPT	36.20	**73.10**	**48.42**	19.19	**66.02**	29.74	**47.11**	**45.77**	**46.43**	59.14	51.60
BioGPT (Luo et al., 2022)	**49.52**	43.25	46.17	**40.00**	39.72	**38.42**	41.70	44.75	40.76	**85.12**	**78.20**

Table 3: Performance on Relation Extraction, Document Classification, and Question Answering datasets.

Model	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
ChatGPT	30.5	12.8	25.4	89.3	28.1	9.8	2.4	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**
BioBART-Base	**61.1**	**48.5**	**59.4**	**94.1**	**46.7**	26.0	44.1	91.8	53.8	36.5	51.3	92.9	30.1	11.3	27.4	89.8	32.9	11.3	29.3	86.1	19.0	7.5	16.8	85.0
BioBART-Large	60.3	48.0	58.7	94.0	46.5	**26.1**	**44.2**	**91.9**	**55.6**	**38.1**	**53.2**	**93.3**	**32.0**	**12.4**	**29.7**	**90.3**	30.6	10.4	27.0	86.1	21.6	9.3	19.2	85.7

Table 4: Performance on Abstractive Summarization datasets. Here, 'R-1', 'R-2', 'R-L' and 'B-S' denote 'ROUGE-1', 'ROUGE-2', 'ROUGE-L', and 'BERTScore', respectively. For BERTScore (Zhang et al., 2019), we use the RoBERTa-Large model (Liu et al., 2019). The results for BioBART models are taken from Yuan et al. (2022a).

lation extraction, ChatGPT led to higher recall scores but much lower precision scores compared to the fine-tuned BioGPT model. This is because ChatGPT tends to generate long and descriptive responses, leading to many inaccurate relation extractions. Though in terms of F1, it outperforms fine-tuned BioGPT in the BC5CDR dataset, it fails to outperform in the KD-DTI dataset. More importantly, it outperforms BioGPT in the DDI dataset in all metrics: Precision, Recall, and F1.

While analyzing the results in different datasets, we observe that in both BC5CDR and DDI datasets where ChatGPT outperforms BioGPT, the training set is small, only 500 and 664 instances, respectively. On the other hand, in the KD-DTI dataset where ChatGPT fails to outperform BioGPT, the training set contains 12000 instances. This gives us a strong indication that even in the biomedical domain, zero-shot ChatGPT can outperform fine-tuned biomedical models in smaller-sized datasets.

We also observe that more descriptive prompts may help ChatGPT to obtain better Precision scores. Contrary to the KD-DTI dataset, we describe the definition of each interaction type in the DDI dataset (see Table 1) where ChatGPT performs the best. To further investigate the effect of prompts in relation extraction, we evaluate the performance in BC5CDR with a new prompt:

i. *Identify the chemical-disease interactions in the passage given below: [PASSAGE].*

We observe that the Precision, Recall, and F1 scores are decreased by 16.07%, 10.3%, and 14.29%, respectively, with this prompt variation.

Document Classification Evaluation: We observe that in the HoC dataset, the zero-shot ChatGPT achieves an F1 score of 59.14, in comparison to its counterpart fine-tuned BioGPT which achieves an F1 score of 85.12. We also investigate the effect of prompt tuning by evaluating with two new prompts that are less descriptive (see Appendix A.2 for more details):

i. *Prompting without explicitly mentioning the name of 10 HoC classes, drops F1 to 38.20.*

ii. *Prompting with the name of each HoC class is given without providing the definition of each class, drops the F1 score to 46.93.*

Question Answering Evaluation: We observe that in the PubMedQA dataset, the zero-shot ChatGPT achieves much lower accuracy than BioGPT (51.60 by ChatGPT in comparison to 78.20 by BioGPT). However, the BioGPT model was fine-tuned on about 270K QA-pairs in various versions of the PubMedQA dataset for this task. While ChatGPT achieves more than 50% accuracy even without any few-shot examples in the prompt.

Summarization Evaluation: We observe that in terms of all ROUGE scores (Lin, 2004), ChatGPT performs much worse than BioBART in datasets that have dedicated training sets, such as iCliniq, HealthCareMagic, and MeQSum. Meanwhile, it performs on par with BioBART in the MEDIQA-QS dataset. More importantly, it 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 training data and ChatGPT achieves comparable or even better performance in these datasets compared to the BioBART model fine-tuned on other related datasets (Yuan et al., 2022a). This further confirms that zero-shot ChatGPT is more useful than domain-specific fine-tuned models in biomedical datasets that lack large training data.

5 Conclusions and Future Work

In this paper, we evaluate ChatGPT on 4 benchmark biomedical tasks to observe that in datasets that have large training data, ChatGPT performs quite poorly in comparison to the fine-tuned models (BioGPT and BioBART), whereas it outperforms fine-tuned models on datasets where the training data size is small. These findings suggest that ChatGPT can be useful in low-resource biomedical tasks. We also observe that ChatGPT is sensitive to prompts, as variations in prompts led to a noticeable difference in results.

Though in this paper, we mostly evaluate ChatGPT on tasks that require it to generate responses by only analyzing the input text, in the future, we will investigate the performance of ChatGPT on more challenging tasks, such as named entity recognition and entity linking (Yadav and Bethard, 2018; Yan et al., 2021; Yuan et al., 2022b; Laskar et al., 2022a,b,c), as well as problems in information retrieval (Huang et al., 2005; Huang and Hu, 2009; Yin et al., 2010; Laskar et al., 2020, 2022d). We will also explore the ethical implications (bias or privacy concerns) of using ChatGPT in the biomedical domain.

Limitations

Since the training datasets of ChatGPT are unknown, some data used for evaluation may or may not exist during the training phase of ChatGPT Also, a new version called the GPT-4 model has been released that may ensure higher accuracy. Nonetheless, GPT-4 is very costly to use, around 60x more expensive than ChatGPT. Meanwhile, even using the paid ChatGPT Plus[2] subscription, it is available for just a limited use (allows evaluation of only 25 samples in 3 hours). Another limitation of this research is that the results mentioned in this paper for ChatGPT may not be reproducible, as ChatGPT may generate different responses for the same input prompt. Although the experimental results may change over time, this work will still give a concrete direction for future research using ChatGPT like large language models in the biomedical domain.

Ethics Statement

The paper evaluates ChatGPT on 4 benchmark biomedical tasks that require ChatGPT to generate a response based on the information provided in the input text. Thus, no data or prompt was provided as input that could lead to ChatGPT generating any responses that pose any ethical or privacy concerns. This evaluation is only done in some academic datasets that already have gold labels available and so it does not create any concerns like humans relying on ChatGPT responses for sensitive issues, such as disease diagnosis. Since this paper only evaluates the performance of ChatGPT and investigates its effectiveness and limitations, conducting this evaluation does not lead to any unwanted biases. Only the publicly available academic datasets are used that did not require any licensing. Thus, no personally identifiable information has been used.

[2] https://openai.com/blog/chatgpt-plus, last accessed June 4th, 2023

Acknowledgements

We would like to thank all the anonymous reviewers for their detailed review comments. This work is done at York University and supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada and the York Research Chairs (YRC) program.

References

Asma Ben Abacha and Dina Demner-Fushman. 2019. On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2228–2234.

Asma Ben Abacha, Yassine M'rabet, Yuhao Zhang, Chaitanya Shivade, Curtis Langlotz, and Dina Demner-Fushman. 2021. Overview of the mediqa 2021 shared task on summarization in the medical domain. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 74–85.

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

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

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

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. 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)*, pages 3615–3620.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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*, pages 4171–4186.

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

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of biomedical informatics*, 46(5):914–920.

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

Xiangji Huang and Qinmin Hu. 2009. 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*, pages 307–314.

Xiangji Huang, Ming Zhong, and Luo Si. 2005. York university at TREC 2005: Genomics track. In *Proceedings of the Fourteenth Text REtrieval Conference, TREC 2005, Gaithersburg, Maryland, USA, November 15-18, 2005*, volume 500-266 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

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

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2022. Ammu: a survey of transformer-based biomedical pretrained language models. *Journal of biomedical informatics*, 126:103982.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Xiangji Huang. 2023. A systematic study and comprehensive evaluation of chatgpt on benchmark datasets. *arXiv preprint arXiv:2305.18486*.

Md Tahmid Rahman Laskar, Cheng Chen, Xue-yong Fu, and Shashi Bhushan Tn. 2022a. Improving named entity recognition in telephone conversations via effective active learning with human in the loop. In *Proceedings of the Fourth Workshop on Data Science with Human-in-the-Loop (Language Advances)*, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Md Tahmid Rahman Laskar, Cheng Chen, Jonathan Johnston, Xue-Yong Fu, Shashi Bhushan TN, and Simon Corston-Oliver. 2022b. 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*, pages 3363–3367.

Md Tahmid Rahman Laskar, Cheng Chen, Aliaksandr Martsinovich, Jonathan Johnston, Xue-Yong Fu, Shashi Bhushan Tn, and Simon Corston-Oliver. 2022c. 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*, pages 344–352, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Md Tahmid Rahman Laskar, Enamul Hoque, and Jimmy Xiangji Huang. 2022d. Domain adaptation with pre-trained transformers for query-focused abstractive text summarization. *Computational Linguistics*, 48(2):279–320.

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

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

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

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6).

Khalil Mrini, Franck Dernoncourt, Seunghyun Yoon, Trung Bui, Walter Chang, Emilia Farcas, and Ndapandula Nakashole. 2021. 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)*, pages 1505–1515.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. 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*, pages 58–65.

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

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

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Max Savery, Asma Ben Abacha, Soumya Gayen, and Dina Demner-Fushman. 2020. Question-driven summarization of answers to consumer health questions. *Scientific Data*, 7(1):1–9.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Vikas Yadav and Steven Bethard. 2018. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158.

Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang, and Xipeng Qiu. 2021. A unified generative framework for various ner subtasks. 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)*, pages 5808–5822.

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

Xiaoshi Yin, Jimmy Xiangji Huang, Xiaofeng Zhou, and Zhoujun Li. 2010. 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*, pages 901–902.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022a. Biobart: Pretraining and evaluation of a biomedical generative language model. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 97–109.

Hongyi Yuan, Zheng Yuan, and Sheng Yu. 2022b. 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*, pages 4038–4048.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. 2023. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *arXiv preprint arXiv:2302.09419*.

A Appendix

A.1 Evaluating ChatGPT on Different Tasks

Since ChatGPT generated responses can be lengthy, and may sometimes contain unnecessary information while not in a specific format, especially in tasks that may have multiple answers (e.g., Relation Extraction), it could be quite difficult to automatically evaluate its performance in such tasks by comparing with the gold labels by using just an evaluation script. Thus, for some datasets and tasks, we manually evaluate the ChatGPT generated responses by ourselves and compare them with the gold labels. Below we describe our evaluation approach for different tasks:

- **Relation Extraction:** The authors manually evaluated the ChatGPT generated responses in this task by comparing them with the gold labels. To ensure the reproducibility of our evaluation, we will release the ChatGPT generated responses.

- **Document Classification:** We created an evaluation script and identifies if the gold label (one of the 10 HoC classes) is present in the ChatGPT generated response. For fair evaluation, we lowercase each character in both the gold label and the ChatGPT generated response. Our evaluation script will be made publicly available to ensure reproducibility of our findings.

- **Question Answering:** Similar to Document Classification, we also evaluated using an evaluation script that compares the gold label and the ChatGPT generated response (here, we also convert each character to lowercase). The evaluation script will also be made public.

- **Abstractive Summarization**: We used the HuggingFace's Evaluate[3] library (Wolf et al., 2020) to calculate the ROUGE scores and the BERTScore for the Abstractive Summarization task evaluation.

[3] https://huggingface.co/docs/evaluate/index

A.2 Effects of Prompt Variation

We investigate the effects of prompt tuning in the HoC dataset by evaluating the performance of ChatGPT based on the following prompt variations:

- Prompting with explicitly defining the 10 HoC classes achieves an F1 score of 59.14 (see Row 1 in Table 5).

- Prompting without explicitly mentioning the name of 10 HoC classes, drops F1 to 38.20 (see Row 2 in Table 5).

- 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 5).

Our findings demonstrate that more descriptive prompts yield better results.

A.3 Sample ChatGPT Generated Responses

Some sample prompts with the ChatGPT generated responses for Relation Extraction, Document Classification, and Question Answering tasks are given in Table 6 and for the Abstractive Summarization task are given in Table 7.

#	Prompt	F1
1.	The 10 hallmark cancer taxonomy with their definitions MeQSum 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. (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 from existing 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) Cellular energetics: 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 evolution of more aggressive and drug-resistant cancer cells. Classify the following sentence in one of the above 10 hallmark cancer taxonomy. If cannot be classified, answer as "empty": [SENTENCE]	59.14
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

Table 5: Effects of Prompt Variation in the HoC dataset for the Document Classification Task.

Dataset	Type	Prompt	ChatGPT 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.	Drug: Famotidine Drug-induced disease: Delirium Side effect: Increased risk of delirium in elderly patients	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'): Analogue inhibitors by modifying oseltamivir based on the crystal neuraminidase structure for treating drug-resistant H5N1 virus. The worldwide spread of H5N1 avian influenza and the increasing reports about its resistance to the existing drugs have made a priority for the development of the new anti-influenza molecules. The crystal structure of H5N1 avian influenza neuraminidase reported recently by Russell et al. [R.J. Russell, L.F. Haire, D.J. Stevens, P.J. Collins, Y. P. Lin, G.M. Blackburn, A.J. Hay, S.J. Gamblin, J.J. Skehel, The structure of H5N1 avian influenza neuraminidase suggests new opportunities for drug design, Nature 443 (2006) 45-49] have provided new opportunities for drug design in this regard. It is revealed through the structure that the active sites of the group-1 neuraminidases, which contain the N1 subtype, have a very different three-dimensional structure from those of group-2 neuraminidases. The key difference is in the 150-loop cavity adjacent to the conserved active site in neuraminidase. Based on these findings and by modifying oseltamivir, six analog inhibitors were proposed as candidates for developing inhibitors against H5N1 virus, particularly against the oseltamivir-resistant H5N1 virus strain.	Drug-target interactions: - Oseltamivir and neuraminidase - Analog inhibitors and neuraminidase Interaction type: Inhibitor	drug: Oseltamivir target: Neuraminidase interaction: inhibitor
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.	1. Leukine and lithium - effect 2. Leukine and corticosteroids - effect 3. Lithium and Leukine - advice 4. Corticosteroids and Leukine - advice	drug: Leukine target: lithium interaction: effect drug: Leukine target: corticosteroids interaction: effect
HoC	Document Classification	The 10 hallmark 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. (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 from existing 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) Cellular energetics: 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 evolution of more aggressive and drug-resistant cancer cells. Classify the following sentence in one of the above 10 hallmark cancer taxonomy. If cannot be classified, answer as "empty": Heterogeneity in DNA damage within the cell population was observed as a function of radiation dose.	genomic instability and mutation.	genomic instability and mutation.
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.

Table 6: Sample prompts with ChatGPT generated responses for Relation Extraction, Document Classification, and Question Answering Tasks.

Dataset	Type	Prompt	ChatGPT Response	Gold Summary
iCliniq	Dialog 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
HealthCare Magic	Dialog 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 don't brush and floss regularly. Bacteria that build up in your mouth and between your teeth produce the bad odor. Other problems in your mouth, such as gum disease, dry mouth or cavities, may also cause it. Sinusitis or problems with your 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 your 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 your 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 your pacemaker.Most appliances in your home are safe to be around. This includes your 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 your 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 your 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 your body as your pacemaker. When using your cell phone, hold it to your ear on the opposite side of your body. Be careful around metal detectors and security wands. Handheld security wands may interfere with your pacemaker. Show your 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. Your pacemaker may set off alarms. After any operation, have your provider check your 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 your 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 your arm for 2 to 3 weeks. Raise your arm above your shoulder for several weeks. When you leave the hospital, you will be given a card to keep in your 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 your card.	It is important to ask your 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 your 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 your cell phone in a pocket or to the ear on the same side as your pacemaker.

Table 7: Sample prompts with ChatGPT generated responses for Abstractive Summarization tasks.