BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining

Original Paper by

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, Tie-Yan Liu

Published in

Briefings in Bioinformatics, September-2022

https://doi.org/10.48550/arXiv.2210.10341

Report By

Bhautikkumar Lukhi

392232 - Advanced AI in Biomedicine (graded)

5th February, 2025

Contents

1 Introduction

The exponential growth of biomedical literature poses significant challenges for researchers worldwide, creating an urgent need for automated tools to extract and interpret knowledge. Web-based search tools, such as PubMed[3], Semantic Scholar[4], arXiv[1], and online bibliographic archives have been developed over the last decades. In 2021, PubMed averaged approximately 2.5 million queries per day[27], indicating its significant role in the biomedical research community. The PubMed[3] database, currently contains more than 36 million entries, and around 1.5 million new items are added each year as of April, 2023(i.e., more than two papers per minute), a rate that itself increases, making PubMed's growth exponential[24].

Biomedical text mining tasks such as named entity recognition (NER), relation extraction, and question answering (QA) are critical for applications including drug discovery, clinical therapy, and pathology research[23, 17]. For instance, NER facilitates identifying biomedical concepts like drug names and diseases[32], while relation extraction enables mapping interactions such as drug-drug or protein-disease associations. These tasks form the foundation of knowledge discovery in the biomedical domain.

Pre-training models have demonstrated their powerful capability in natural language processing (NLP). On the GLUE benchmark[36], a widely used benchmark for natural language understanding, pre-training based methods outperform non-pre-training methods by a large margin. These models are first pre-trained on large scale corpora collected from the Web via self-supervised learning task and then fine-tuned on specific downstream tasks. BERT-like models are widely used for sequence classification and labeling, requiring complete document encoding. In contrast, GPT-like models excel in tasks like abstract and knowledge triplet generation. In the biomedical domain, models like BioBERT[19] and PubMedBERT[10] have achieved notable success in understanding and processing text for classification and labeling tasks. However, their capabilities are limited when it comes to generative tasks, which are increasingly important for applications such as text generation and extracting complex knowledge representations.

BioGPT bridges this gap with a generative pre-trained Transformer tailored for biomedical applications. Trained on 15 million PubMed abstracts, it achieves state-of-the-art results in six key biomedical NLP tasks, including relation extraction, question answering, document classification, and text generation. By refining target sequence formats and task-specific prompts, BioGPT advances biomedical NLP, streamlining knowledge discovery in the domain.

This report provides an in-depth analysis of the BioGPT model, examining its architecture, training methodology, performance on downstream tasks, and its transformative implications for biomedical research.

2 Background

2.1 Pre-trained Language Models in Biomedicine

In Figure 1 we can see chronological overview of large language models (LLMs) and their variants in biomedicine from 2019 to 2024. The timeline showcases the progression of unimodal and multimodal models, emphasizing significant advancements in various architectures such as LLAMA[34], GPT[29], BERT[8], CLIP[28] and others. BioGPT was first to successfully adopt the GPT model in the biomedical domain and show state-of-the-art results.

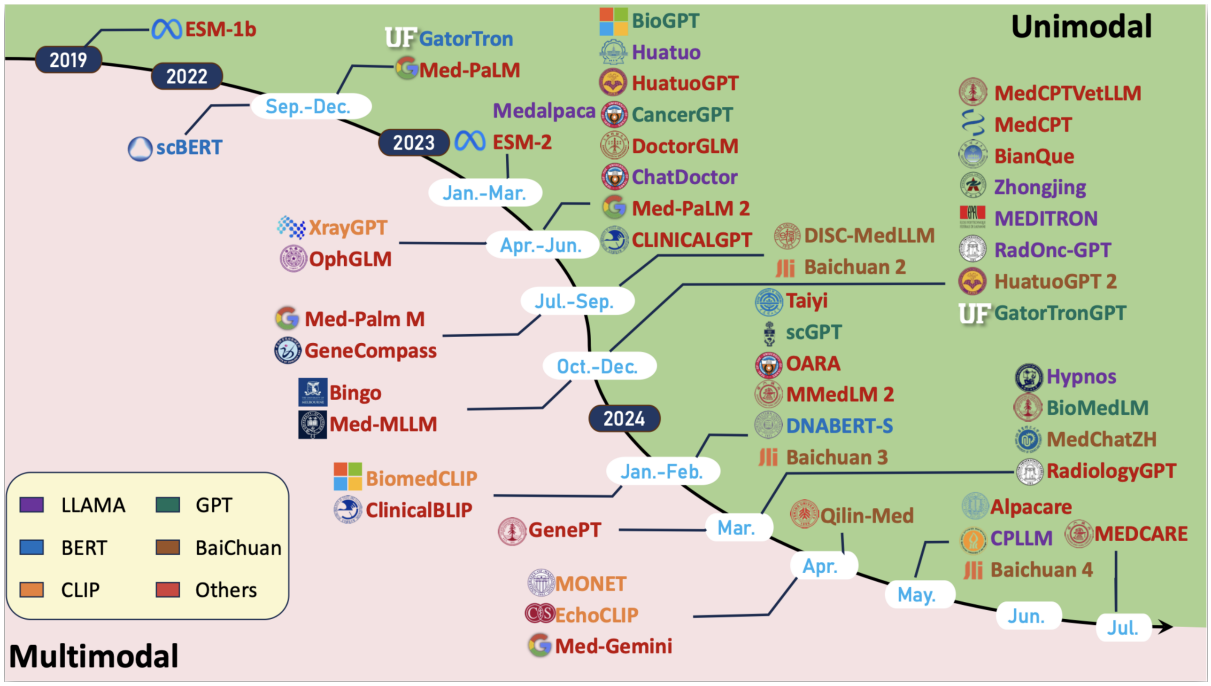

Figure 1: A Survey for Language Models in Biomedicine

The rapid growth and diversification of LLM research in biomedicine are further evidenced by the trends shown in Figure 2. A temporal analysis of LLM research papers in biomedical fields from 2018 to 2024 reveals an increase in publications, with a surge beginning in 2021 (Fig. 2a). This trend underscores the growing interest and investment in applying LLMs to biomedical challenges, reflecting both the technological advancements and the recognition of LLMs' potential to address healthcare and research needs. The distribution of these research papers across various biomedical fields highlights 'medicine' and 'neuroscience' as the dominant areas of focus (Fig. 2b). This distribution demonstrates the broad applicability of LLMs across different medical specialties and research domains, while also indicating potential areas for future expansion and development[37].

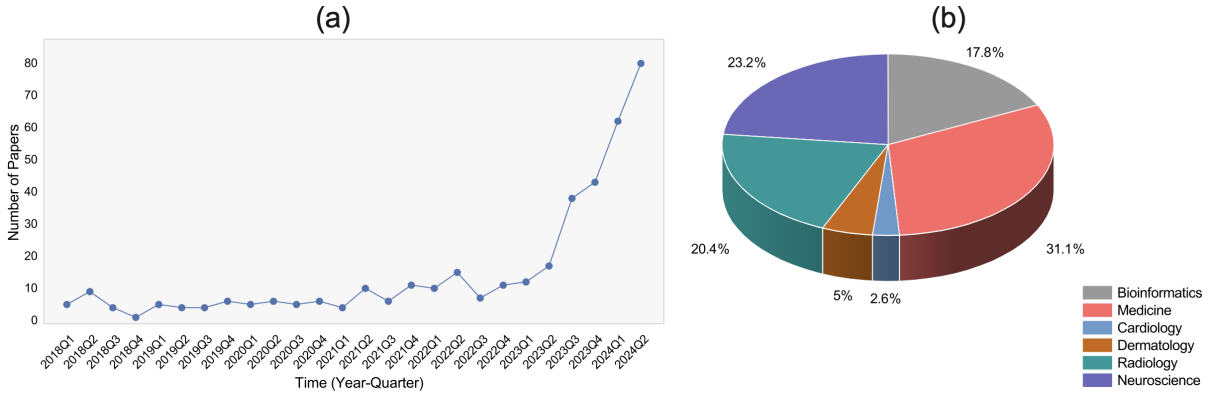

Figure 2: Trends and distribution of LLM research papers in biomedical fields

2.2 Domain-Specific Pretraining: Challenges and Advances

Applying general-domain pre-trained models like BERT to biomedicine often leads to subpar performance due to domain mismatch. To address this, models such as BioBERT[19] and ClinicalBERT[14] continue pretraining on biomedical-specific corpora like PubMed abstracts and full-text articles. However, these models retain the original BERT vocabulary, which is suboptimal for biomedical text.

To overcome this limitation, PubMedBERT[10] pretrains from scratch on 14 million PubMed abstracts, creating a vocabulary tailored to the biomedical domain. Similarly, an ELECTRA-based model[22] pretrains from scratch on 28 million biomedical texts, yielding further improvements. These efforts demonstrate that domain-specific pretraining significantly enhances understanding tasks like NER and relation extraction.

However, generative tasks remain underexplored. While GPT models excel at generation, studies show they perform poorly on biomedical tasks[6, 12]. For instance, DARE[25] pretrained GPT on only 0.5 million PubMed abstracts, limiting its use to data augmentation. This highlights the need for a biomedical-specific generative Transformer. BioGPT fills this gap by pretraining on 15 million PubMed abstracts, enabling superior performance in biomedical text generation and mining.

Pre-trained language models have demonstrated remarkable success in NLP. These models can be broadly categorized into:

- **BERT-like models:** Primarily focused on language understanding tasks using bidirectional contextual embeddings.
- **GPT-like models:** Designed for language generation tasks using auto-regressive language modeling.

Models like BioBERT and PubMedBERT[10] adapt BERT for biomedical tasks but fail to address generative needs. GPT models, while generative, perform poorly on biomedical tasks due to domain shift[26].

Model	Strengths	Limitations
BERT	Versatile for general NLP tasks	Limited domain understanding
BioBERT	Excellent for NER and relation extraction	Shared vocabulary with general domain
BlueBERT	Good biomedical language understanding	Shared vocabulary with general domain
SciBERT	Adaptable across science domains	Out-of-domain for specific biomedical terms
PubMedBERT	Optimized vocabulary for biomedical text	Encoder-only architecture
ELECTRAMed	Highly efficient for classification tasks	Encoder-only architecture

Table 1: Comparison of Pre-trained Biomedical Models

3 Architecture and Training

3.1 Model Design

BioGPT is based on the GPT-2_{medium} architecture[30], as illustrated in Figure 3. This architecture features 16 attention heads[35], 24 layers, and a hidden size of 1024, totaling 355M parameters. However, BioGPT's distinctiveness lies not in its core architecture but in the adaptations made for its biomedical applications:

- **Domain-Specific Vocabulary:** BioGPT uses a customized tokenizer and vocabulary tailored to biomedical texts. This ensures accurate handling of complex biomedical terms, abbreviations, and domain-specific entities. This adjustment, along with modifications in the embedding and output projection layers, results in BioGPT's parameter count being slightly lower at 347M.
- **Pretraining on Biomedical Texts:** Instead of relying on generic web text, BioGPT was pretrained from scratch on biomedical literature found on PubMed[3]. This allows the model to excel in tasks specific to the biomedical domain, such as relation extraction and scientific text generation.

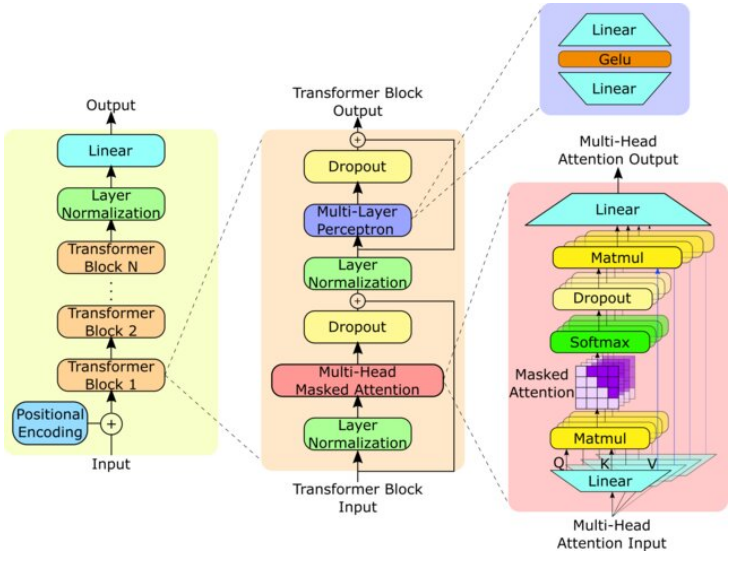

Figure 3: GPT-2 architecture overview

These modifications and training methodologies enable BioGPT to specialize in generating accurate, domain-specific biomedical text.

3.1.1 Multi-head Attention Layer

The multi-head attention is the fundamental element of both Transformer and BioGPT. Three linear transformations are performed on the input to yield the value V, the key K, and the query Q. The result is then computed as follows:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$

$$\text{where head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

Where the projections are parameter matrices:

$$W_i^Q \in \mathbf{R}^{d_{\text{model}} \times d_k}, \quad W_i^K \in \mathbf{R}^{d_{\text{model}} \times d_k}, \quad W_i^V \in \mathbf{R}^{d_{\text{model}} \times d_v}, \quad W^O \in \mathbf{R}^{hd_v \times d_{\text{model}}}.$$

The output of the multi-head attention layer is then fed into a feed-forward layer to construct a Transformer layer.

3.2 Training Criteria

Similar to [29, 30]. BioGPT is trained using the conventional language modeling task. Let $D = \{x_i\}_i$ denote the collection of sequences, and sequence x_i is made up of n_i tokens, i.e., $x_i = (s_1, s_2, \ldots, s_{n_i})$. The training objective is to minimize the negative log-likelihood:

$$\min -\frac{1}{|D|} \sum_{i=1}^{|D|} \sum_{j=1}^{n_i} \log P(s_j \mid s_{j-1}, s_{j-2}, \ldots, s_1). \tag{1}$$

3.3 Pre-training Dataset

The quantity, quality, and domain of the dataset are critical for pre-training language models. Training solely on in-domain data from start to finish is crucial for a particular domain, as Gu et al. [11] point out. Thus, BioGPT is pre-trained from scratch using the gathered data to stay true to the domain. While selecting abstracts from PubMed only in-domain biomedical texts were included and abstracts with missing information were filtered out.

3.4 Vocabulary Development

Gu et al.[11] highlight the critical role of in-domain vocabulary for domain-specific language models. BioGPT learns the language using it's gathered in-domain corpus rather

than the GPT-2 vocabulary. To learn the vocabulary, authors specifically divide the corpus's words into word segments using byte pair encoding (BPE)[33] with the use of the fastBPE5 BPE implementation. Final size of the learned vocabulary is 42384.

3.5 Training Setup and Hardware

BioGPT was pre-trained and evaluated on four biomedical NLP tasks using six datasets. The tasks included end-to-end relation extraction on BC5CDR[20], KD-DTI[13], and DDI[31]; question answering on PubMedQA[15]; document classification on HOC[5]; and text generation on a self-created dataset.

- **Hardware:** 8 NVIDIA V100[2] GPUs were used.
- **Batch Size:** 524,288 tokens (calculated as $1024 \times 8 \times 64$).
- **Optimizer:** Adam[18] was employed with a peak learning rate of 2×10^{-4}, including a warm-up phase of 20,000 steps. The learning rate followed an inverse square root decay schedule after the warm-up phase.
- **Training Steps:** A total of 200,000 iterations were conducted.

4 Fine-tuning for Downstream Tasks

The pre-trained BioGPT was further adapted for three downstream tasks: End-to-end relation extraction, Question Answering (QA), and Document Classification. To maintain consistency with the pre-trained format, task labels were converted into natural language sequences rather than using structured formats with special tokens. This approach ensures smoother semantic alignment between pre-training and fine-tuning tasks.

4.1 End-to-End Relation Extraction

The objective of this task is to extract relational triplets $\langle \text{head entity}_i, \text{tail entity}_i, \text{relation}_i \rangle$ from a given input sequence. To simplify the task, relational triplets were reformulated into natural language sentences in three formats:

1. **Subject-Verb-Object(SVO):** e.g., *Dextropropoxyphene inhibits mu-type opioid receptor.*
2. **Is-of Form:** e.g., *Dextropropoxyphene is the inhibitor of mu-type opioid receptor.*
3. **Rel-is Form:** e.g., *The relation between Dextropropoxyphene and mu-type opioid receptor is inhibitor.*

For documents with multiple triplets, the sentences were sorted by their order of appearance and concatenated using semicolons. These natural language sentences can be converted back into triplets using regular expressions, with flexibility for task-specific formatting.

4.2 Question Answering

In this task, the goal was to answer questions based on a given reference context. The input sequence was constructed by appending descriptive tags, such as *question:* and *context:*, to the respective text. The output sequence followed the format: *The answer to the question given the context is [label].* For example:

- **Input:** *question: What is the cause of disease X? context: Disease X is caused by gene Y.*
- **Output:** *The answer to the question given the context is gene Y.*

4.3 Document Classification

The objective of document classification was to assign a category to a given document. The output sequence was generated in the format: *The type of this document is [label].* For example: *The type of this document is genomic instability and mutation.*

This natural language-based approach to label representation not only aligns with BioGPT's pre-training on textual data but also allows for easy adaptability to various downstream tasks.

4.4 Prompt-based Fine-tuning

The labels were formatted as target sequences, and a prompt-based approach was adopted to facilitate fine-tuning and inference with BioGPT. A straightforward method of concatenating the source and target sequences was deemed impractical, as it did not provide clear task-specific guidance to the model during inference.

To address this, prompt-based technique[21] was employed as seen in the Figure 4, which involve appending task-specific instructions to the input to guide the model in generating task-relevant outputs. While manually designed discrete prompts (hard prompts) have been explored in previous works, their effectiveness is highly dependent on the prompt design, which can be labor-intensive and inconsistent across tasks. Instead, authors primarily utilized soft prompts through prefix-tuning. This method incorporates continuous embeddings, or virtual tokens, that are appended as prompts. These embeddings are randomly initialized and fine-tuned end-to-end on downstream tasks, ensuring they are task-specific. Unlike conventional prefix-tuning approaches, where virtual tokens

are appended at the beginning of the source input, BioGPT appends the virtual tokens between the source and target sequences.

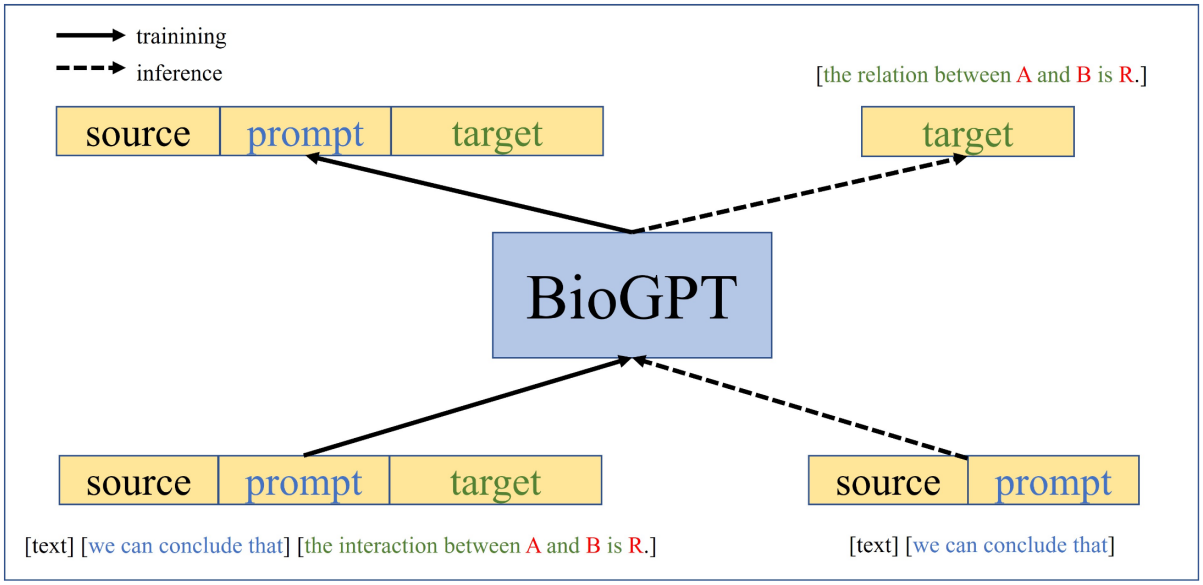

Figure 4: Framework of BioGPT when adapting to downstream tasks

During inference, the source text and prompt serve as the prefix for the language model, guiding it to generate the corresponding target output. This method ensures seamless adaptation of the pre-trained BioGPT to various downstream tasks.

5 Evaluation and Results

5.1 End-to-End Relation Extraction

BioGPT was evaluated for end-to-end relation extraction, where the model directly generates relational triplets from input text. Its performance was compared with REBEL[9] and seq2rel[7], both based on sequence-to-sequence models, as well as GLRE[38], a pipeline-based method requiring NER annotations.

5.1.1 BC5CDR Dataset

It is a binary chemical-disease relation dataset with 500 documents each for training, validation, and testing[20]. For evaluation models were fine-tuned for 100 epochs using a learning rate of 10^{-5}, and the averaged checkpoints of the last 5 epochs were used. The micro-F1 score was the primary metric. The detailed results are presented in Table 2 below, comparing BioGPT's performance with other state-of-the-art models. BioGPT achieved the highest F1 score of 44.98%, outperforming GLRE[38] and REBEL[9] and seq2rel[7]. Even when seq2rel was trained on both the training and validation sets, BioGPT outperformed it.

Model	Precision	Recall	F1
GLRE (gt+pred) [38]	34.82	18.29	23.99
GLRE (pred+pred [38])	23.00	4.88	8.05
GPT-2 [30]	43.92	32.55	37.39
REBEL [9]	34.28	39.49	36.70
REBELpt [9]	40.94	21.20	27.94
seq2rel [7]†	43.50	37.50	40.20
BioGPT	**49.44**	**41.28**	**44.98**
BioGPT†	**49.52**	**43.25**	**46.17**

Table 2: Results on BC5CDR binary chemical-diseass relation extraction task.

5.1.2 KD-DTI Dataset

It is a drug-target interaction dataset with 12k/1k/1.3k documents for training, validation, and testing respectively[13]. Similarly models were fine-tuned for 30 epochs using Adam[18] optimizer with a peak learning rate of 10^{-5}. BioGPT achieved 38.42% F1, surpassing Transformer + PubMedBERT-attn [13] by 14.23%, GPT-2_{medium}[30] by 9.97% and REBEL [9] by 8.03%. BioGPT also surpassed REBEL_{pt}[9], which uses additional pretraining, by 5.1%.

Model	Precision	Recall	F1
Transformer + PubMedBERT -attn [13]	25.35	24.14	24.19
GPT-2_{medium} [30]	30.53	27.87	28.45
REBEL [9]	32.36	29.58	30.39
REBEL_{pt} [9]	35.73	32.61	33.32
BioGPT	**40.00**	**39.72**	**38.42**

Table 3: Results on KD-DTI drug-target interaction extraction task.

5.1.3 DDI Extraction 2013 Corpus

It is a Drug-drug interaction dataset with 664/50/191 files for training, validation, and testing[31]. Models fine-tuned for 100 epochs with a learning rate of 10^{-4}. BioGPT achieved 40.76% F1, with 16.08% improvement over GPT-2_{medium}[30], 12.49% improvement over REBEL[9] and slightly surpassing REBEL_{pt}[9] (40.56%).

5.2 Question Answering

PubMedQA[15] is a biomedical question answering dataset. Each sample comprises a PubMed abstract containing a question, a reference context, a long answer, and a yes/no/maybe label as the answer.They use the original train/validation/test split with 450, 50, and 500 samples, respectively, noted as PQA-L in [15] for evaluation.

Model	Precision	Recall	F1
GPT-2_{medium} [30]	23.39	31.93	24.68
REBEL [9]	35.36	28.64	28.27
REBELpt [9]	46.59	39.60	40.56
BioGPT	41.70	**44.75**	**40.76**

Table 4: Results on DDI drug-drug-interaction extraction task.

Model	Accuracy
PubMedBERT [10]	55.8
BioELECTRa [16]	64.2
BioLinkBERT$_{\text{base}}$ [39]	70.2
BioLinkBERT$_{\text{large}}$ [39]	72.2
BioGPT	**78.2**

Table 5: Results on PubMedQA question answering task.

Table 5 presents the classification accuracy for the reasoning-required setting from [15]. BioGPT achieves 78.2% accuracy, outperforming BioLinkBERT[39] by 6.0% and setting a new state-of-the-art.

5.3 Document Classification

The Hallmarks of Cancer (HoC) corpus[5] consists of 1580 PubMed abstracts annotated at the sentence level with ten hallmark labels. Following the training/test split in [15], continuous embedding of length 1 was used as the soft prompt and format labels into target sequences. The GPT-2_{medium}[30] and BioGPT were fine-tuned for 20,000 steps using a peak learning rate of 10^{-5} and 1000 warm-up steps.

As Table 6 shows BioGPT achieves an F1 score of 85.12%, surpassing GPT-2_{medium}[30] by 3.28% and outperforming domain-specific models such as BioBERT[19] and PubMedBERT[10] .

Model	F1
BioBERT [19]	81.54
PubMedBERT [10]	82.32
PubMedBERT$_{\text{large}}$ [10]	82.70
BioLinkBERT$_{\text{base}}$ [39]	84.35
GPT-2_{medium} [30]	81.84
BioGPT	**85.12**

Table 6: Results on HoC Document Classification Task.

5.4 Text Generation

Pre-trained models like GPT [29], GPT-2 [30], and GPT-3 [6] exhibit impressive text generation capabilities. The authors evaluate the biomedical text generation ability of BioGPT against GPT-2_{medium} [30] using entities from the KD-DTI [13] test set as prefixes (e.g., drug and target names). The generated texts are assessed for fluency and relevance.

Input	**Model**	**Summary Of Generated Text**
Bicalutamide	GPT-2	Describes cellular proliferation in *C. elegans*.
	BioGPT	Describes clinical use as an AR antagonist for prostate cancer treatment.
Janus kinase 3	GPT-2	Links to glucose metabolism and reduction in muscle protein breakdown.
	BioGPT	Defines its role in cell proliferation, differentiation, and angiogenesis.
Xylazine	GPT-2	Mentions as a component of "bath salts" linked to deaths.
	BioGPT	Defines its role as a sedative/analgesic in veterinary medicine and its effects on cardiovascular and CNS systems.

Table 7: Comparison of GPT-2 and BioGPT on Common Biomedical Keywords

BioGPT consistently produces precise and professional biomedical descriptions, while GPT-2 struggles with uncommon or domain-specific terms. For example, in Table 7 above BioGPT accurately describes the clinical use of Bicalutamide as an androgen receptor (AR) antagonist for prostate cancer treatment. In contrast, GPT-2 generates irrelevant content about cellular proliferation in *C. elegans*, which is unrelated to the actual biomedical context.

Similarly, as shown in Table 8 below, BioGPT effectively explains COVID-19 as a pandemic caused by SARS-CoV-2, providing coherent and accurate descriptions. On the other hand, GPT-2 generates incoherent links and unrelated descriptions due to its pre-training data cutoff of 2021.

These comparisons underscore BioGPT's superior domain-specific understanding and text generation capabilities, especially in biomedical and COVID-19-related contexts. Overall, BioGPT exhibits superior text generation in the biomedical domain, producing meaningful and fluent text even for highly specific or novel biomedical entities.

Input	Model	Summary Of Generated Text
COVID-19	GPT-2	Incoherent links and unrelated descriptions.
	BioGPT	Defines COVID-19 as a pandemic caused by SARS-CoV-2.
Treatment of COVID-19	GPT-2	Mentions irrelevant effects on dopamine systems.
	BioGPT	Discusses remdesivir as FDA-approved treatment.
Omicron variants	GPT-2	Unclear explanation about strain misidentification.
	BioGPT	Explains pathogenicity and role in severe infections.

Table 8: Comparison of GPT-2 and BioGPT on COVID-19 Related Topics

6 Scaling to Larger Size

To further enhance BioGPT's capabilities, the authors developed BioGPT-Large, based on the GPT-2 XL architecture with 1.5 billion parameters. Fine-tuning and evaluation on downstream tasks demonstrate notable improvements in performance. Key results include:

Task	Metric	Performance
BC5CDR	F1	50.12
KD-DTI	F1	38.39
DDI	F1	44.89
PubMedQA	Accuracy	81.0
HoC	F1	84.40

Table 9: Performance of BioGPT-Large fine-tuned on downstream tasks.

This scaling highlights BioGPT's potential for even greater impact in biomedical natural language processing with larger models.

7 Conclusion

BioGPT is a specialized generative pre-trained Transformer model developed to address challenges in biomedical text generation and mining. Leveraging the architecture of GPT-2 [30] and fine-tuned on 15 million PubMed[3] abstracts, BioGPT is designed to understand and generate domain-specific biomedical text. Its superior performance across various tasks such as document classification, relation extraction, and question answering illustrates its effectiveness in processing and generating accurate, meaningful biomedical content.

To further scale its impact, $\text{BioGPT}_{\text{Large}}$, based on the GPT-2_{XL}[30] architecture with

1.5 billion parameters, has shown promising results across multiple downstream tasks. This scaling demonstrates $\text{BioGPT}_{\text{Large}}$'s potential for even greater accuracy and robustness when applied to increasingly complex biomedical datasets. Notably, $\text{BioGPT}_{\text{Large}}$ has shown significant improvement in tasks like relation extraction and drug-target interaction prediction, confirming the model's capacity to scale effectively without compromising performance. This scalability positions $\text{BioGPT}_{\text{Large}}$ as an increasingly powerful tool in Bio-medical Natural Language Procession(BioNLP), with the potential for broader applications across drug discovery, medical research, and healthcare. When compared to GPT-2[30], BioGPT also demonstrates remarkable capabilities in generating domain-relevant biomedical text, such as correctly describing the clinical uses of drugs or explaining medical concepts like COVID-19. These advancements underscore BioGPT's exceptional understanding of biomedical language, not only in terms of syntax and grammar but also in its ability to contextualize and generate highly specific and technical content.

In conclusion, BioGPT represents a significant milestone in the field of biomedical natural language processing and knowledge discovery. Its ability to generate high-quality, contextually accurate biomedical text makes it an invaluable resource for researchers and practitioners alike. The further development of BioGPT, particularly through the creation of larger model, undoubtedly leads to more advanced applications and deeper insights into the biomedical domain, facilitating the growth of knowledge and innovation in medical science and healthcare.

References

[1] arXiv. https://arxiv.org.

[2] NVIDIA V100. https://www.nvidia.com/en-au/data-center/v100/.

[3] PubMed - National Library of Biomedicine. https://pubmed.ncbi.nlm.nih.gov.

[4] Semantic Scholar. https://www.semanticscholar.org.

[5] Simon Baker, Ilona Silins, Yufan Guo, Imran Ali, Johan Högberg, Ulla Stenius, and Anna Korhonen. Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinformatics*, 32 3:432–40, 2016.

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Ma teusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.

[7] Xi Chen, Julien Cumin, Fano Ramparany, and Dominique Vaufreydaz. Generative resident separation and multi-label classification for multi-person activity recognition. *2024 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, pages 1–6, 2024.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 2019.

[9] Zhaolin Gao, Jonathan D. Chang, Wenhao Zhan, Owen Oertell, Gokul Swamy, Kianté Brantley, Thorsten Joachims, J. Andrew Bagnell, Jason D. Lee, and Wen Sun. Rebel: Reinforcement learning via regressing relative rewards. *ArXiv*, abs/2404.16767, 2024.

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

[11] Yu Gu, Robert Tinn, Hao Cheng, Michael R. Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language

model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3:1 – 23, 2020.

[12] Bernal Jimenez Gutierrez, Nikolas McNeal, Clay Washington, You Chen, Lang Li, Huan Sun, and Yu Su. Thinking about gpt-3 in-context learning for biomedical ie? think again. 2022.

[13] Yutai Hou, Yingce Xia, Lijun Wu, Shufang Xie, Yang Fan, Jinhua Zhu, Tao Qin, and Tie-Yan Liu. Discovering drug–target interaction knowledge from biomedical literature. *Bioinformatics*, 38(22):5100–5107, 10 2022.

[14] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *ArXiv*, abs/1904.05342, 2019.

[15] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *ArXiv*, abs/1909.06146, 2019.

[16] Kamal Raj Kanakarajan, Bhuvana Kundumani, and Malaikannan Sankarasubbu. Bioelectra:pretrained biomedical text encoder using discriminators. In *Workshop on Biomedical Natural Language Processing*, 2021.

[17] Halil Kilicoglu. Biomedical text mining for research rigor and integrity: Tasks, challenges, directions. *bioRxiv*, 2017.

[18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

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

[20] 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. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database: The Journal of Biological Databases and Curation*, 2016, 2016.

[21] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *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 4582–4597, 2021.

[22] Giacomo Miolo, Giulio Mantoan, and Carlotta Orsenigo. Electramed: a new pre-trained language representation model for biomedical nlp. *ArXiv*, abs/2104.09585, 2021.

[23] Anthony N. Nguyen, John O'Dwyer, Thanh Vu, Penelope M. Webb, Sharon E Johnatty, and Amanda B. Spurdle. Generating high-quality data abstractions from scanned clinical records: text-mining-assisted extraction of endometrial carcinoma pathology features as proof of principle. *BMJ Open*, 10, 2020.

[24] J. Novoa, M. Chagoyen, C. Benito, F. J. Moreno, and F. Pazos. Pmidigest: Interactive review of large collections of pubmed entries to distill relevant information. *Genes*, 14(4):942, 2023.

[25] Yannis Papanikolaou and Andrea Pierleoni. Dare: Data augmented relation extraction with gpt-2. *ArXiv*, abs/2004.13845, 2020.

[26] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *BioNLP@ACL*, 2019.

[27] Zhiyong Lu Qiao Jin, Robert Leaman. Pubmed and beyond: biomedical literature search in the age of artificial intelligence. *eBioMedicine*, 100:104988, 2023.

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. 2021.

[29] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018.

[30] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[31] Karsten Boye Rasmussen and Daniel W. Gillman. Ddi and semantic web. 2015.

[32] Shuaib Raza, Deepak John Reji, F. Shajan, and Syed Raza Bashir. Large-scale application of named entity recognition to biomedicine and epidemiology. *PLOS Digital Health*, 1, 2022.

[33] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *ArXiv*, abs/1508.07909, 2015.

[34] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971, 2023.

[35] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017.

[36] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[37] Chong Wang, Mengyao Li, He Junjun, Zhongruo Wang, Erfan Darzi, Zan Chen, Jin Ye, Tianbin Li, Yanzhou Su, Jing Ke, Kaili Qu, Shuxin Li, Yi Yu, Pietro Lio, Tianyun Wang, Yu Guang Wang, and Yiqing Shen. A survey for large language models in biomedicine. August 2024.

[38] D. Wang, Wei Hu, Ermei Cao, and Weijian Sun. Global-to-local neural networks for document-level relation extraction. *ArXiv*, abs/2009.10359, 2020.

[39] Michihiro Yasunaga, Jure Leskovec, and Percy Liang. Linkbert: Pretraining language models with document links. In *Annual Meeting of the Association for Computational Linguistics*, 2022.