

BioMedLM: A 2.7B Parameter Language Model Trained On Biomedical Text

Elliot Bolton^{1†}, Abhinav Venigalla², Michihiro Yasunaga¹, David Hall¹, Betty Xiong¹,
Tony Lee¹, Roxana Daneshjou¹, Jonathan Frankle²,

Percy Liang¹, Michael Carbin², and Christopher D. Manning^{1†}

¹Stanford University

²DataBricks

[†]Correspondence: elliottbolton@stanford.edu, manning@stanford.edu

March 2024

Abstract

Models such as GPT-4 and Med-PaLM 2 have demonstrated impressive performance on a wide variety of biomedical NLP tasks. However, these models have hundreds of billions of parameters, are computationally expensive to run, require users to send their input data over the internet, and are trained on unknown data sources. Can smaller, more targeted models compete? To address this question, we build and release BioMedLM, a 2.7 billion parameter GPT-style autoregressive model trained exclusively on PubMed abstracts and full articles. When fine-tuned, BioMedLM can produce strong multiple-choice biomedical question-answering results competitive with much larger models, such as achieving a score of 57.3% on MedMCQA (dev) and 69.0% on the MMLU Medical Genetics exam. BioMedLM can also be fine-tuned to produce useful answers to patient questions on medical topics. This demonstrates that smaller models can potentially serve as transparent, privacy-preserving, economical and environmentally friendly foundations for particular NLP applications, such as in biomedicine. The model is available on the Hugging Face Hub: <https://huggingface.co/stanford-crfm/BioMedLM>.

1 Introduction

Large language models such as OpenAI’s GPT-4 have become the dominant technology in modern natural language processing (Liu et al., 2023; Zhao et al., 2023). Trained on large corpora to predict the next token and refined with human feedback (Brown et al., 2020; Ouyang et al., 2022; Ziegler et al., 2020), these models develop impressive capabilities in areas such as summarization and question-answering (Zhang et al., 2023; Goyal et al., 2023; Karpukhin et al., 2020). While the focus has been on these models’ performance when responding to general English prompts, it is clear there is potential for specialist models to impact biomedical research and healthcare (Arora and Arora, 2023; Shah et al., 2023; Thirunavukarasu et al., 2023). Such applications include information retrieval and summarization from the ever-expanding biomedical literature (Wang et al., 2021; Yang, 2020), clinical information such as physician notes in electronic health records, and radiology reports (Murray et al., 2021; Feblowitz et al., 2011; Zhang et al., 2018). Improving domain-specific language models will help accelerate biomedical discovery, drive down healthcare costs, and improve patient care.

Large, general models like GPT-4 and Med-PaLM 2 have set new standards for performance on question-answering and information extraction (Kung et al., 2022; Singhal et al., 2023a,b), but there are several drawbacks to these models.

They are costly to train and utilize. Compute for training and inference of large language models have increased 10- to 100-fold since 2015 (Sevilla et al., 2022), translating to extremely high financial and

environmental costs (Sharir et al., 2020; Patterson et al., 2021). Organizations must pay expensive API fees which may be beyond their budget as well.

Due to their immense size and corporate secrecy, these models must be run on giant compute clusters and accessed remotely over the internet. This eliminates the possibility of on-device inference and requires the transmission of sensitive data. The need to access these models via corporate APIs (Cobbe et al., 2023), raises issues around third-party access of personal identifiable information (PII) (Lukas et al., 2023). This conflicts with the medical need for patient data privacy, outlined by the Health Insurance Portability and Accountability Act (HIPAA) (Marks and Haupt, 2023; Meskó and Topol, 2023).

The closed nature of these models create many complications as well. Any organization relying on these models is ultimately relying on a major tech company. If the tech company ceases to serve the model for any reason, an organization could immediately find any service it built on top of the model deactivated.

One of the most prominent drawbacks of this closed nature is that the training data used for corporate models is a closely guarded secret. From a business perspective, this creates a risk to an organization using these models. For example, if an AI model provider has trained their model on data that violates copyright, the model may unpredictably become unavailable. From a practitioner perspective, the lack of transparency around model training data adds uncertainty to the quality and reliability of model responses. And from a research perspective, not knowing the constitution of model training data inhibits the ability to study the relationship between training data and downstream task performance.

Another consequence of the closed nature of these models is the inability to further fine-tune them for specific tasks. This is especially meaningful in the biomedical context, where models trained on general English data can benefit from further training on specialized tasks (Singhal et al., 2023a,b).

In summary, these expenses and challenges can make language model technology inaccessible or non-viable to biomedical and/or healthcare organizations with limited resources and strict privacy requirements.

A promising direction is to build a smaller model and focus training on biomedical text, to achieve good domain-specific performance, and provide hope for a smaller solution. However, models like PubMedBERT (Gu et al., 2021), SciBERT (Beltagy et al., 2019) and BioBERT (Lee et al., 2019) are much smaller by current standards. This presents an opportunity for a small-medium domain-specific model with strong performance. To further investigate the value of domain specialization, and the potential of smaller models, we introduce BioMedLM, a 2.7 billion parameter biomedical language model trained on PubMed abstracts and full articles, and made publicly available in December 2022. It is a GPT-2 style (Radford et al., 2019) autoregressive model (Vaswani et al., 2017) with biomedical domain-specific tokenization, and trained exclusively on PubMed abstracts and full articles (Gao et al., 2020).

BioMedLM can achieve strong results on multiple-choice biomedical question answering, which are competitive with substantially larger models. This includes scores of 57.4% on MedMCQA (dev) (Pal et al., 2022) and 70.0% on MMLU Medical Genetics (Hendrycks et al., 2021). Against a general-English model baseline, GPT-Neo 2.7B (Gao et al., 2020) (architecturally similar model to BioMedLM with a nearly identical parameter count), BioMedLM demonstrates an improvement in accuracy across three tasks: BioASQ (Tsatsaronis et al., 2015), PubMedQA (Jin et al., 2019) and MedQA (Jin et al., 2021). We also demonstrate question-answering requiring text generation capabilities; BioMedLM can produce useful multi-sentence answers to questions on medical topics from the HealthSearchQA question dataset (Singhal et al., 2023a), such as “What are the best ways to treat plantar fasciitis?”.

BioMedLM is designed to address the drawbacks mentioned above. Its small size allow for it to be comfortably fine-tuned on a single A100 GPU and for inference to be run on a laptop. Organizations utilizing this model can serve it internally and never send private data out of their internal networks. Its training data is fully documented, allowing practitioners and researchers insight into model performance. And the model’s open nature allows anyone to download it and fine-tune it as they see

fit.

As competition intensifies in AI, corporations are increasingly keeping the details of their models’ architectures and training secret. Thus, it is vital for organizations to release models to the open source community so researchers can learn about the capabilities of language models and the trade-offs of different architecture and training design choices (Kapoor et al., 2024). We hope BioMedLM can spread knowledge about how to build LLMs and contribute to the body of knowledge around training domain-specific models and applying language models to biomedical NLP tasks. By releasing BioMedLM, we provide a performant smaller model which can be of use for biomedical tasks of language analysis and production.

2 Related Work

There are two broad types of models that have been deployed for biomedical question-answering NLP tasks. Large, general English models that have been adapted for the biomedical domain, like Med-PaLM 2 (Singhal et al., 2023b), are the state of the art on multiple-choice question-answering and multi-sentence answering. Small, domain-targeted models like DRAGON (Yasunaga et al., 2022a) have shown impressive performance on multiple choice and text classification tasks for their size.

GPT-Neo 2.7B. GPT-Neo 2.7B is a GPT-style model with 2.7B parameters (Black et al., 2021) trained on the Pile, a diverse English corpus 825 GB in size (Gao et al., 2020). The Pile contains multiple sub-corpora, including general content from Common Crawl and specialized content like PubMed, Github, and FreeLaw. This model serves as an important baseline for comparison with our domain-specific models; because it is the same size as our model and has a similar architecture, we can better understand the impact of pre-training corpus composition on downstream task performance.

PubMedBERT. PubMedBERT is a BERT model trained on PubMed abstracts and articles that uses a biomedical-oriented tokenizer (Gu et al., 2021). This model showed substantial improvement on biomedical NLP tasks over similar models trained on biomedical NLP tasks such as NER, relation extraction, and question-answering, demonstrating the potential of exclusively training on PubMed.

BioLinkBERT and DRAGON. BioLinkBERT and DRAGON are BERT-style models trained on PubMed (Yasunaga et al., 2022b,a). They utilize augmented architectures and richer data sources to produce highly impressive performance on BLURB and biomedical question-answering (Gu et al., 2021). BioLinkBERT exploits the link structure present in PubMed, while DRAGON builds on BioLinkBERT to leverage information from common biomedical knowledge graphs. With their high quality, domain-specific data, both models are able to compete with substantially larger models, despite only having 350-million parameters.

Galactica. Galactica is a 120 billion parameter language model trained on scientific text that utilizes a tokenizer specialized for scientific text (Taylor et al., 2022). It achieved impressive zero-shot performance on scientific tasks, including biomedical QA tasks, e.g., PubMedQA (Jin et al., 2019), further demonstrating the value of training on specialized text and scale.

BioGPT. BioGPT is a GPT-style model trained on PubMed concurrent to this work (Luo et al., 2022). This work shows the capabilities of a biomedical GPT-style model, including state-of-the-art results on PubMedQA when using artificial and unlabeled data and strong results on a variety of biomedical NLP tasks. While the BioGPT paper focuses more on relation extraction and summarization, we emphasize question-answering tasks.

Flan-PaLM and Med-PaLM. Flan-PaLM and Med-PaLM are fine-tuned versions of the 500 billion parameter PaLM language model (Chowdhery et al., 2022; Singhal et al., 2023a,b). Flan-PaLM set the standard for MedQA and passed the USMLE without being specifically trained to perform on biomedical tasks, showing the capabilities of large-scale models trained on general language.

ChatGPT and GPT-4. Kung et al. (2022) evaluates ChatGPT (GPT-3.5) on a collection of publicly available USMLE questions and finds that ChatGPT can generally achieve passing scores without additional fine-tuning on specialized data. Nori et al. (2023a) assesses the capabilities of GPT-4 on USMLE and MultiMedQA benchmarks, and demonstrates that it exceeds performance of the previous GPT-3.5 and Med-PaLM. Nori et al. (2023b) presents Medprompt, based on a composition of several prompting strategies, which elevates GPT-4 to better performance than Med-PaLM 2. While it is believed that scale can lead to significant progress in biomedical NLP, it is also difficult to assess train-test leakage in closed models.

Fully Open Language Models And Datasets. A thriving open source community is promoting full openness in model and dataset construction (i.e., training data, training process, and architecture are publicly known). Some of the first prominent models in this class were released by EleutherAI, including GPT-J (Wang and Komatsuzaki, 2021) and GPT-NeoX-20B (Black et al., 2022) which were both trained on the Pile. Together’s RedPajama (Together, 2023b) project, which aims to produce high quality open-source English datasets, has released 1 trillion and 30 trillion token datasets that have been used to train numerous open source models, including the 3 and 7 billion parameter INCITE models(Together, 2023a). Similarly the Allen Institute for AI has released OLMo (Groeneveld et al., 2024) trained on the open Dolma (Soldaini et al., 2024) training set.

3 Model Design And Training

3.1 Model Architecture

BioMedLM is an autoregressive, decoder-only Transformer (Vaswani et al., 2017) with 2.7 billion parameters. It is similar in architecture to GPT-2 (Radford et al., 2019) with the settings in Table 1.

Parameter	Setting
Hidden Size	2560
Heads	20
Layers	32
Vocab Size	28896
Sequence Length	1024

Table 1: Model Architecture Settings

The model uses learned absolute positional embeddings for each position in the sequence (Vaswani et al., 2017).

3.2 BioMedLM Tokenizer

It has been shown across several domains that domain-specific tokenizers help (Sachidananda et al., 2021), including the biomedical domain (Gu et al., 2021; Yasunaga et al., 2022b,a), science in general (Taylor et al., 2022) and law (Zheng et al., 2021). Utilizing tokenizers customized for their domains allows information for important terms to be stored in corresponding embeddings, rather than split across multiple sub-term tokens.

BioMedLM uses a custom Byte-Pair Encoding (BPE) tokenizer (Sennrich et al., 2016) that was trained on PubMed abstracts. The tokenizer was trained with the Hugging Face Tokenizers library (Huggingface, 2019) with the settings in Table 2.

Parameter	Setting
Vocab Size	28896
Min Frequency	2
Add Prefix Space	False
Trim Offsets	True

Table 2: Tokenizer Training Settings

Using this custom tokenizer results in many common biomedical terms being represented by a single token. For instance, Table 3 shows a sampling of biomedical terms that are split by the traditional GPT-2 tokenizer, but are not split by our BioMedLM tokenizer.

BioMedLM Tokenization	GPT2 Tokenization
chromatography	chrom/atography
cytotoxicity	cyt/ot/oxicity
ECG	EC/G
GATA	G/ATA
Immunohistochemistry	Imm/un/oh/ist/ochemistry
myocardium	my/ocard/ium
nanoparticles	nanop/articles
photosynthesis	photos/ynthesis
probiotic	probiotic
thrombin	th/rom/bin

Table 3: Examples of BioMedLM and GPT2 Tokenizations of Biomedical Terms

Consider a term like “thrombin”. The standard Hugging Face GPT-2 tokenizer splits this token into “th”, “rom”, and “bin”. With the standard tokenizer, information about thrombin is spread across these three tokens and shared with completely unrelated words that have small sub-word overlap. Compounding the issue, these sub-word tokens do not correspond to meaningful units the way tokens such as “bio” or “photo” might. The BioMedLM tokenizer preserves this word as a single token, without storing information in disparate tokens.

This improved treatment of domain-specific terminology translates to improved downstream task performance. We ran multiple experiments at the 125 million parameter scale to test out possible design choices. Experiments with different tokenizers but otherwise identical settings showed a meaningful improvement on the MedQA task over 5 random seeds (see Table 4).

Model	MedQA Test Accuracy
BioMedLM-125m (GPT-2 tokenizer)	33.05
BioMedLM-125m (BioMedLM tokenizer)	34.98

Table 4: Tokenizer Impact on MedQA Performance

3.3 Pre-training

BioMedLM was trained on the subparts of The Pile (as of November 2022) containing PubMed abstracts and full articles (Gao et al., 2020). There were 34.6 billion tokens in the training corpus. The training run performed 8.67 passes through the data.

The model was trained on MosaicML Cloud, a platform designed for large workloads like training LLMs. The code used during pre-training utilized Flash Attention (Dao et al., 2022), an algorithm for accelerating and reducing the memory requirements of calculating attention. This was a crucial implementation detail that allowed us to pre-train a model of BioMedLM’s scale on our available compute resources. We used Hugging Face’s GPT-2 model code (Radford et al., 2019), the Composer training library (MosaicML, 2021) and PyTorch FSDP (Paszke et al., 2019), employing multi-node training across 128 40GB Nvidia A100 GPUs. The total training run was completed in 6.25 days.

The model was trained with batch size 1024 and sequence length 1024 (~ 1 million tokens per batch) for 300 billion tokens using Decoupled AdamW (Loshchilov and Hutter, 2019) to minimize the standard cross entropy loss of the subsequent token. The settings used for training are summarized below in Table 5.

Parameter	Setting
Tokens Per Batch	1048576
Learning Rate	1.6e-4
Scheduler	Cosine w/linear warmup (100 batches)
Epsilon	1e-8
Betas	[0.9, 0.95]
Weight Decay	1.6e-5

Table 5: Pre-training Settings

Initial trial runs at the 1.5 billion scale suffered from consistent training loss divergences. These runs used fp16/fp32 mixed precision (using NVIDIA’s AMP library) similar to the Mistral project (Karamcheti et al., 2021), generally using fp16 precision but upcasting to fp32 when computing attention. The divergence issues were resolved when we switched to using bf16 (Wang and Kanwar, 2019).

The final training process of the 2.7 billion parameter model did not suffer from any divergences. During the final run, parameters and optimizer states were stored in fp32, gradient communication was done in fp32, and training was done in bf16 (see Table 6).

Parameter	Library	Setting
Compute Precision	Composer	bf16
Parameter Storage	FSDP	fp32
Optimizer Storage	FSDP	fp32
Gradient Communication	FSDP	fp32

Table 6: Mixed Precision Settings

As we were preparing the training run, we were unsure of the benefits of training out to 300B tokens for language model perplexity and downstream task performance. While most models of this scale (e.g., GPT-Neo 2.7B (Gao et al., 2020)) are trained to 300-400B tokens, the datasets those models use are vastly larger than PubMed. For instance, The Pile is 8x the size of its PubMed subcorpora.

Fortunately, we did continue to see steady perplexity improvements on the validation and training sets for the entirety of training (Figure 1), and preliminary experiments showed improved downstream task performance as we trained out to the full 300B tokens. Consistent with concurrent work showing that long training is useful for small models (De Vries, 2023; Touvron et al., 2023), our takeaway is that it was indeed worth it to train for the full 300B tokens, despite having dramatically more passes through the data than comparable models (e.g., Karamcheti et al., 2021).

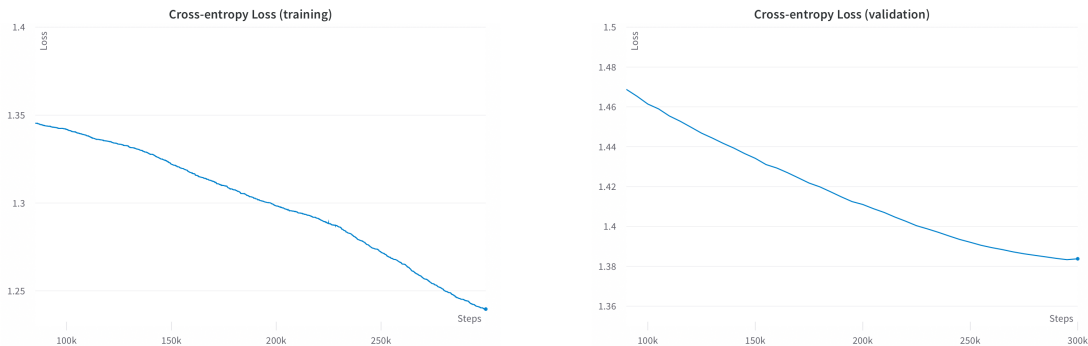


Figure 1: Train and Validation Loss after 100k Batches

3.4 Fine-tuning

Fine-tuning for Question-Answering Tasks. We fine-tune the pre-trained BioMedLM for downstream question-answering tasks. Different tasks have different output formats. For MedMCQA and MedQA, the prompt contains a passage that ends with a question, from which the model is asked to pick the correct option. For MMLU, we fine-tuned a multiple-choice model on the MedMCQA and MedQA training data and evaluated it on the selection of MMLU exams related to biomedicine used in previous work (Singhal et al., 2023a).

For generative models, it is common to supply the question text with answer options as the input context and record the generated answer from the model’s response. The model’s performance can be improved by including few-shot examples in the input context or directly fine-tuning the model to respond to multiple choice questions with the appropriate answer.

Our experiments demonstrated multiple accuracy points of improvement by using an architecture specialized for multiple-choice exams. For each question, the question context is concatenated with each answer option. Each question context and ending is run through the transformer, and the hidden state at the end of the full sequence with the proposed ending is run through a linear classifier which creates scores for each ending which are used to select the answer. This multiple-choice-specific architecture was used for our MedQA, MedMCQA, and MMLU results.

For PubMedQA and BioASQ, the context and question were concatenated, and the hidden state for a special token at the end of the concatenated sequence was fed to a linear classifier which produced scores for “yes/no/maybe” or “yes/no”. Effectively the questions were treated as sequence classification examples for these two tasks. A small set of experiments directly training BioMedLM to generate “yes”, “no”, or “maybe” for PubMedQA yielded similar results, so it is not clear if the sequence classification architecture provided any performance boost.

Examples for each question type can be found in the supplementary material.

Standard Hugging Face code for these specializations can be found on the GitHub accompanying this paper: <https://github.com/stanford-crfm/BioMedLM>.

Optimal Format for PubMedQA and BioASQ. Since BioMedLM is a unidirectional model, it requires extra care during fine-tuning for the BioASQ and PubMedQA tasks. Unlike MedQA which simply contains a passage that ends with a question, BioASQ and PubMedQA come in a structured format that separates the question and context passage, leaving it to the user to arrange these components into prompts for the language model. In our initial experiments we placed the question at the beginning and followed it with the context. We were able to see significant performance gains on these tasks when we introduced special tokens and reorganized the components into the following format:

[Context Token] “Text of context ...” [Question Token] “Text of question ...” [Answer Token]

Fine-tuning for Long-Form Question-Answer Generation. To evaluate our model’s generative capabilities, we fine-tuned it on a collection of medical knowledge question/answer pairs directed towards patients. First, we collected over 53,000 questions from publicly available sources on the web. A typical example would be a question like “What are the best ways to treat plantar fasciitis?” and the corresponding answer a medical doctor would give.

4 Biomedical Question-Answering Results

With a trained model in hand, we began exploring its capabilities in the biomedical QA space. We focused on 5 standard biomedical NLP QA tasks: MedMCQA (Pal et al., 2022), MedQA (Jin et al., 2021), MMLU (Hendrycks et al., 2021), PubMedQA (Jin et al., 2019), and BioASQ (Tsatsaronis et al., 2015), as well as summarization of consumer health questions (Ben Abacha and Demner-Fushman, 2019). Our model was able to produce results competitive with substantially larger models, and was often able to surpass the bi-directional models with augmented architectures trained on richer biomedical data. Examples of all these question-answering datasets are shown in the Appendix.

4.1 MedMCQA

This dataset contains 182822/4183/6150 train/dev/test questions drawn from AIIMS PG and NEET PG questions found on the web and in books (Pal et al., 2022). There is a wide variety of questions covering topics ranging from clinical questions to fundamental biochemistry, with each question having four multiple-choice options. Results are presented in Table 7.

Model	Params	Method	MedMCQA Test
			Accuracy
GPT-4	–	few-shot	72.4
Flan-PaLM	540B	few-shot	57.6
BioMedLM	2.7B	fine-tune	57.3
Galactica	120B	zero-shot	52.9
GPT-3.5	175B	few-shot	51.0

Table 7: MedMCQA Performance of Various Systems

4.2 MedQA

This dataset contains 10178/1272/1273 train/dev/test questions drawn from USMLE questions found on the web (Kung et al., 2022). A standard question presents a medical scenario that a physician should be able to answer, and four options. Results are presented in Table 8.

Model	Params	Model Openness	Method	MedQA Test Accuracy
Med-PaLM 2	–	closed	few-shot	85.4
GPT-4	–	closed	few-shot	81.4
Flan-PaLM	540B	closed	few-shot	67.2
BioMedLM (MedMCQA data + classifier)	2.7B	fully open	fine-tune	54.7
GPT-3.5	175B	closed	few-shot	53.6
BioMedLM (classifier)	2.7B	fully open	fine-tune	50.3
DRAGON	360M	fully open	fine-tune	47.5
BioLinkBERT	340M	fully open	fine-tune	45.1
Galactica	120B	open weights	zero-shot	44.4
GPT-Neo 2.7B	2.7B	fully open	fine-tune	37.7

Table 8: MedQA Performance of Various Systems

4.3 MMLU

The MMLU collection of questions covers a wide variety of subject areas, from biology and mathematics to history and philosophy (Hendrycks et al., 2021). The tests also range over different academic levels, from high school to professional. Several of the tests are relevant to the biomedical domain. Results are presented in Table 9.

Model	Params	Method	Clinical Knowledge	Professional Medicine	College Biology	Medical Genetics
GPT-4	–	few-shot	86.4	93.8	93.8	92.0
Flan-PaLM	540B	few-shot	80.4	83.8	88.9	75.0
GPT 3.5	175B	zero-shot	69.8	70.2	72.2	70.0
Galactica	120B	zero-shot	59.2	59.6	68.8	70.0
BioMedLM	2.7B	fine-tune	59.6	63.1	60.7	69.0

Table 9: MMLU Performance of Various Systems

4.4 PubMedQA

This dataset contains 450/50/500 train/dev/test questions constructed from combining example PubMed article abstracts with questions derived by altering the paper title and answers related to the abstract (Jin et al., 2019). Each question can be answered yes/no/maybe. Results are presented in Table 10.

Model	Params	Method	PubMedQA Test Accuracy
BioGPT (w/extra data)	1.5B	fine-tune	81.0
Flan-PaLM	540B	few-shot	79.0
Galactica	120B	zero-shot	77.6
GPT-4	-	zero-shot	75.20
BioMedLM	2.7B	fine-tune	74.4
DRAGON	360M	fine-tune	73.4
BioLinkBERT	340M	fine-tune	72.4
GPT-Neo 2.7B	2.7B	fine-tune	66.1

Table 10: PubMedQA Performance of Various Systems

4.5 BioASQ

This task is similar to PubMedQA, containing 670/75/140 train/dev/test questions constructed from combining example biomedical passages and questions relevant to the passage (Tsatsaronis et al., 2015). Each question can be answered yes/no. Results are presented in Table 11.

Model	Params	Method	BioASQ
			Test Accuracy
DRAGON	360M	fine-tune	96.4
BioMedLM	2.7B	fine-tune	95.7
BioLinkBERT	340M	fine-tune	94.9
Galactica	120B	zero-shot	94.3
GPT-Neo 2.7B	2.7B	fine-tune	67.1

Table 11: BioASQ Performance of Various Systems

5 Free Response Question Answering

While multiple-choice question-answering can demonstrate a model’s medical knowledge, it is not straight forward to turn a multiple-choice answering system into a practical tool. Ideally, one should be able to direct medical questions to the model and get back multi-sentence generated answers, as if consulting an expert colleague.

To test our model’s ability to produce paragraph level responses, we fine-tuned it on health question answer pairs derived from publicly available data on the web. We constructed a training set based on FAQ pages and converting subsections of Wikipedia into a question answer format. Content from FAQ pages was directly added to the training set, with each question on the page being added as one training example. Wikipedia articles were used to generate more examples. For instance, the “Signs and symptom” section of the “Plantar fasciitis” Wikipedia page was converted to the question “What are the signs and symptoms of plantar fasciitis?” with the corresponding section text used as the answer. Thousands more question answer pairs were derived from Wikipedia in this manner.

It is important to note that BioMedLM needs to be fine-tuned on an instruction data set like this to enable it to answer questions like this. The base BioMedLM model was trained to replicate text in PubMed articles, so it is not well suited for question response.

In Table 12, we demonstrate our system’s performance on some example questions. We are planning a more formal evaluation of generation quality in future work.

6 Analysis

6.1 BioMedLM Performance Across Different Biomedical Tasks

At this point, systems with hundreds of billions of parameters have become dominant on some medical tasks (e.g. MedQA), and are rivaling expert human performance. However, in all question-answer tasks (MedMCQA, MedQA, MMLU, PubMedQA and BioASQ), BioMedLM is able to achieve comparable results to larger models.

For MedMCQA, the substantially larger training data set size of 182822 allows BioMedLM, despite being 200x smaller, to match the performance of Flan-PaLM (57.38 vs. 57.6). For MedQA, BioMedLM (54.1) even outperforms the larger Galactica model (44.4) by a significant margin. For MMLU, BioMedLM is able to come close to matching Flan-PaLM on the Medical Genetics topic (70.0 and

Question	BioMedLM Answer
What does vitamin D do?	<p>Vitamin D plays an important role in maintaining your bones and supporting your immune system. It's also involved in many other body processes.</p> <p>Vitamin D is a fat-soluble vitamin. Your body needs to get this vitamin from your diet and/or your sun exposure. You can get vitamin D from foods that contain vitamin D, such as fatty fish, eggs and fortified milk. You can also get vitamin D from sun exposure.</p> <p>Your body needs to get enough vitamin D to function properly. Vitamin D deficiency can cause rickets in children and osteomalacia in adults. Vitamin D deficiency can also cause muscle weakness and an increased risk of falls and bone fractures.</p> <p>Vitamin D deficiency is a common condition. In the United States, about 1 in 10 people have vitamin D deficiency.</p>
What are good treatments for arthritis?	<p>There are many treatments for arthritis. Your healthcare provider will recommend the treatment that's right for you based on your symptoms and the type of arthritis you have. Treatments for arthritis include:</p> <p>Medications: Your healthcare provider may prescribe medications to reduce pain and inflammation. Common medications include:</p> <p>Nonsteroidal anti-inflammatory drugs (NSAIDs).</p> <p>Corticosteroids.</p> <p>Pain relievers.</p> <p>Physical therapy: Physical therapy can help you manage arthritis symptoms. Your healthcare provider may recommend exercises to strengthen muscles and improve flexibility.</p> <p>Surgery: You may need surgery if your arthritis is severe or you don't respond to other treatments. Your healthcare provider may recommend joint replacement surgery or osteotomy.</p> <p>Complications of arthritis treatments You may experience side effects from arthritis treatments. Your healthcare provider will discuss the benefits and risks of your treatment options with you.</p>
What are the common causes of lower back pain?	<p>The most common causes of lower back pain include:</p> <p>Degenerative disk disease: Damage to the disks in your lower back causes your disks to wear down. This wear and tear can lead to lower back pain.</p> <p>Spinal stenosis: A narrowing of your spinal canal can cause pressure on your nerves and lower back pain.</p> <p>Scoliosis: A curved spine can cause pain in your lower back.</p> <p>Herniated disk: A bulge in one of your disks causes pain.</p> <p>Spinal infection: Infections in your spine can cause pain and other symptoms.</p> <p>Pregnancy: Your body is more sensitive during pregnancy, which can cause back pain.</p> <p>Trauma: A fall, car accident or other injury can cause pain in your lower back.</p>
What are the best ways to unclog your ears?	<p>There are several ways to unclog your ears. Your healthcare provider will recommend the method that's right for you.</p> <p>Ear drops: Your healthcare provider will apply ear drops to your ears. The drops may contain a steroid or antibiotic.</p> <p>Ear tubes: Your healthcare provider will surgically insert a small tube into your ear canal. The tube drains fluid and helps drain ear infections.</p> <p>Surgery: Your healthcare provider may perform ear surgery to remove excess ear tissue or open up your ear canal.</p>
How can I stop diarrhea?	<p>The best way to stop diarrhea is to treat the cause. If you have an infection, your healthcare provider may prescribe antibiotics. If you have a food poisoning outbreak, you can go to the emergency room or call your healthcare provider to get help. If you have chronic diarrhea, you may need to change your diet. If you have an autoimmune disease, your healthcare provider may prescribe medication. If you take a medication that causes diarrhea, you may need to stop taking it.</p>

Table 12: BioMedLM Example Responses To Consumer Health Questions

75.0, respectively), and can generally match Galactica on topics with medical emphasis, i.e., Professional Medicine (61.76 and 59.6, respectively). It performs worst on the College Biology topic (57.64), which is likely due to the domain drift between college level biology topics and formal PubMed articles.

For PubMedQA, due to the small training set size, our best system fine-tuned for this task is further away, despite BioMedLM (74.4) being trained on PubMed articles. Recently BioGPT (81.0) has shown great results by being fine-tuned on the extra noisy label and unlabeled data provided with PubMedQA, and we feel a promising future direction would be exploring their 2-phase fine-tuning approach to further improve BioMedLM’s performance. In BioASQ, BioMedLM (95.7) outperforms BioLinkBERT (94.9) and Galactica (94.3).

6.2 Question-Answering with Text Generation

The model’s responses to medical questions are generally accurate. At times the model is clearly delivering vague answers and referring the patient to a medical professional. It is worth noting that several questions get very detailed answers.

One issue that persists is hallucination, especially around numerical values. Both numerical examples in Table 12 appear to be wrong. Basic fact-checking reveals that 35% of people in the US have Vitamin D deficiency rather than 10%. A complete application to provide answers to patients would need additional safeguards to correct incorrect numerical values such as these examples. However, as of now, BioMedLM does have the potential to be a component of such an application.

6.3 Comparison with GPT-Neo 2.7B Baseline

Multiple works have demonstrated that domain-specific training can be beneficial for domain-specific downstream tasks (Zheng et al., 2021). To explore the extent to which pre-training on domain-specific tasks can help, we ran several experiments fine-tuning GPT-Neo 2.7B on biomedical QA tasks. GPT-Neo 2.7B is an architecturally similar model to BioMedLM with a nearly identical parameter count, but trained on general English. Figure 2 shows the relative performance of the two systems on select tasks.

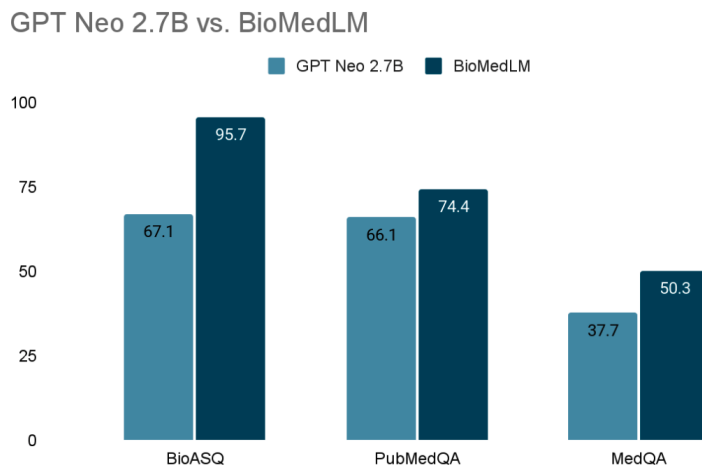


Figure 2: Comparison of GPT-Neo 2.7B and BioMedLM on Select QA Tasks

On each of the three select tasks, BioMedLM substantially outperforms GPT-Neo 2.7B, including a 27% accuracy increase on BioASQ. Given that there can be substantial gains at the 2.7B scale by pre-training on PubMed for a long time, it suggests larger models could also benefit on these types of tasks if pre-trained for longer specifically on PubMed.

7 Usage of BioMedLM since release

Since its release in December 2022, BioMedLM has been evaluated on several biomedical benchmarks. It was evaluated on medical QA tasks (MedQA, PubMedQA) with accuracy 50.3% and 74.4%, respectively (Tian et al., 2023). It has been used to improve the transferability of clinical note section classification models, using data such as discharge summaries, colorectal clinical notes and progress notes (Zhou et al., 2023). In Yun et al. (2023), it was evaluated on text generation of medical systematic reviews after it was given the title of a review article. It has also been evaluated for general medical named entity recognition (Deußer et al., 2023) using bio-entity tasks JNLPBA and NCBI-Disease. BioMedLM performance was state of the art (with micro F1-score of 81.31%), showing that pre-training on a specific domain helps with named entity decoding.

Other benchmarks involve relation extraction and classification, in various biomedical domains. For protein pathways and interactions (Park et al., 2023), tasks include recognizing protein-protein interactions (STRING Task 1 and 2), KEGG pathway recognition (KEGG Task 1 and 2) and evaluating gene regulatory relations (INDRA DB). Interestingly, in STRING Task 2 (classifying existence of any association or interaction between two proteins), BioMedLM demonstrated the most favorable performance (with micro F1-score of 0.643), while larger models, e.g., LLaMA, Alpaca, and BioGPTLarge, exhibited higher rates of false positives, possibly influencing a bias towards positive answers.

In the microbiome field, BioMedLM was fine-tuned to extract microbiome-disease interactions (Karkera et al., 2023). BioMedLM exhibited the best precision (0.822), outperforming GPT-3 (0.810). However, GPT-3 had the best accuracy of 0.814 vs. BioMedLM’s 0.806 and BioGPT’s 0.732, possibly owing to its pre-training corpus of PubMed articles. In genetics, BioMedLM has also been evaluated on a benchmark, GeneTuring test, which focuses on nomenclature, genomic location, functional analysis and sequence alignment (Jin et al., 2023). In the multi-modal space, there have also been efforts to use BioMedLM in vision-language pre-training models for computer-aided diagnosis (Chen et al., 2023); or prefix tuning for visual question answering tasks (van Sonsbeek et al., 2023).

In summary, given BioMedLM’s medium size, as a public, low-cost model, it punches above its weight. There exists a tension between domain-specific knowledge either hindering or enhancing performance, where the former can be explained by over-optimization on medical text corpora, and the latter suggests the benefits of domain-specific information. BioMedLM’s training on domain knowledge in PubMed lends it to better performance in biology-related tasks over clinical tasks. In relation extraction tasks, lack of extraneous out-of-domain data gives it a lower false positive rate and higher precision compared to other models.

8 Conclusion

We present BioMedLM, a 2.7 billion parameter GPT-style model trained on PubMed text. On biomedical question-answering tasks, BioMedLM can outperform bidirectional models with richer data sources and compete with the few-shot performance of models that have orders of magnitude more parameters. BioMedLM can also produce useful generations, delivering multi-sentence answers to medical knowledge questions. Our work demonstrates the potential of medium-sized models trained on domain-specific text.

9 Acknowledgments

We thank DataBricks (originally, MosaicML) for providing the compute for this project.

10 Reproducibility

The pre-trained model is available on the Hugging Face hub:

<https://huggingface.co/stanford-crfm/BioMedLM>

Code used for pre-training and fine-tuning is available on GitHub:

<https://github.com/stanford-crfm/BioMedLM>

References

- Anmol Arora and Ananya Arora. The promise of large language models in health care. *The Lancet*, 401(10377):641, 2023. doi: 10.1016/s0140-6736(23)00216-7.
- Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text, 2019. URL <https://arxiv.org/abs/1903.10676>.
- Asma Ben Abacha and Dina Demner-Fushman. On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2228–2234, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1215. URL <https://aclanthology.org/P19-1215>.
- Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow, March 2021. URL <https://doi.org/10.5281/zenodo.5297715>. If you use this software, please cite it using these metadata.
- Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. Gpt-neox-20b: An open-source autoregressive language model, 2022.
- 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, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL <https://arxiv.org/abs/2005.14165>.
- Qiuhui Chen, Xinyue Hu, Zirui Wang, and Yi Hong. MedBLIP: Bootstrapping language-image pre-training from 3D medical images and texts, 2023. URL <https://arxiv.org/abs/2305.10799>.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. PaLM: Scaling language modeling with pathways, 2022. URL <https://arxiv.org/abs/2204.02311>.
- Jennifer Cobbe, Michael Veale, and Jatinder Singh. Understanding accountability in algorithmic supply chains. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’23*, page 1186–1197, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594073. URL <https://doi.org/10.1145/3593013.3594073>.
- Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness, 2022. URL <https://arxiv.org/abs/2205.14135>.
- Harm De Vries. Go smol or go home, 2023. URL <https://www.harmdevries.com/post/model-size-vs-compute-overhead/>.

- Tobias Deußer, Lars Hillebrand, Christian Bauckhage, and Rafet Sifa. Informed named entity recognition decoding for generative language models, 2023. URL <https://arxiv.org/abs/2308.07791>.
- Joshua C. Feblowitz, Adam Wright, Hardeep Singh, Lipika Samal, and Dean F. Sittig. Summarization of clinical information: A conceptual model. *Journal of Biomedical Informatics*, 44(4):688–699, 2011. doi: 10.1016/j.jbi.2011.03.008.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020. URL <https://arxiv.org/abs/2101.00027>.
- Tanya Goyal, Junyi Jessy Li, and Greg Durrett. News summarization and evaluation in the era of GPT-3, 2023. URL <https://arxiv.org/abs/2209.12356>.
- Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. Olmo: Accelerating the science of language models, 2024. arXiv preprint arXiv:2402.00838.
- 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 Transactions on Computing for Healthcare*, 3(1):1–23, 2021. doi: 10.1145/3458754.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. URL <https://arxiv.org/abs/2009.03300>.
- Huggingface. Huggingface/tokenizers: fast state-of-the-art tokenizers optimized for research and production, 2019. URL <https://github.com/huggingface/tokenizers>.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, Jul 2021. ISSN 2076-3417. doi: 10.3390/app11146421. URL <http://dx.doi.org/10.3390/app11146421>.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering, 2019. URL <https://arxiv.org/abs/1909.06146>.
- Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. GeneGPT: Augmenting large language models with domain tools for improved access to biomedical information, 2023. URL <https://arxiv.org/abs/2304.09667>.
- Sayash Kapoor, Rishi Bommasani, Kevin Klyman, Shayne Longpre, Ashwin Ramaswami, Peter Ci-hon, Aspen Hopkins, Kevin Bankston, Stella Biderman, Miranda Bogen, Rumman Chowdhury, Alex Engler, Peter Henderson, Yacine Jernite, Seth Lazar, Stefano Maffulli, Alondra Nelson, Joelle Pineau, Aviya Skowron, Dawn Song, Victor Storch, Daniel Zhang, Daniel E. Ho, Percy Liang, and Arvind Narayanan. On the societal impact of open foundation models, 2024. URL <https://crfm.stanford.edu/open-fms/paper.pdf>.
- Sidd Karamcheti, Laurel Orr, Jason Bolton, Tianyi Zhang, Karan Goel, Avanika Narayan, Rishi Bommasani, Deepak Narayanan, Tatsunori Hashimoto, Christopher D. Manning, Dan Jurafsky, Christopher Potts, Christopher Ré, and Percy Liang. Mistral — a journey towards reproducible language model training, 2021. URL <https://crfm.stanford.edu/2021/08/26/mistral.html>.

- Nikitha Karkera, Sathwik Acharya, and Sucheendra K. Palaniappan. Leveraging pre-trained language models for mining microbiome-disease relationships. *BMC Bioinformatics*, 24(290), 2023. doi: <https://doi.org/10.1186/s12859-023-05411-z>.
- Vladimir Karpukhin, Barlas Öguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. Dense passage retrieval for open-domain question answering, 2020. URL <https://arxiv.org/abs/2004.04906>.
- Tiffany H. Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, and et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, 2022. doi: 10.1101/2022.12.19.22283643.
- 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 biomedical text mining. *Bioinformatics*, 36(4):1234–1240, Sep 2019. doi: 10.1093/bioinformatics/btz682. URL <https://doi.org/10.1093%2Fbioinformatics%2Fbtz682>.
- Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, and Bao Ge. Summary of ChatGPT/GPT-4 research and perspective towards the future of large language models, 2023. URL <https://arxiv.org/abs/2304.01852>.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL <https://arxiv.org/abs/1711.05101>.
- Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. Analyzing leakage of personally identifiable information in language models, 2023. URL <https://arxiv.org/abs/2302.00539>.
- Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6), 2022. doi: 10.1093/bib/bbac409.
- Mason Marks and Claudia E. Haupt. AI chatbots, health privacy, and challenges to HIPAA compliance. *JAMA*, 330(4):309, 2023. doi: 10.1001/jama.2023.9458.
- Bertalan Meskó and Eric J. Topol. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *npj Digital Medicine*, 6(1), 2023. doi: 10.1038/s41746-023-00873-0.
- MosaicML. Composer. <https://github.com/mosaicml/composer/>, 2021.
- Luke Murray, Divya Gopinath, Monica Agrawal, Steven Horng, David Sontag, and David R Karger. MedKnowts: Unified documentation and information retrieval for electronic health records. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. ACM, October 2021. doi: 10.1145/3472749.3474814. URL <https://doi.org/10.1145%2F3472749.3474814>.
- Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of GPT-4 on medical challenge problems, 2023a. URL <https://arxiv.org/abs/2303.13375>.
- Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine, 2023b. URL <https://arxiv.org/abs/2311.16452>.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL <https://arxiv.org/abs/2203.02155>.

- Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedMCQA : A large-scale multi-subject multi-choice dataset for medical domain question answering, 2022. URL <https://arxiv.org/abs/2203.14371>.
- Gilchan Park, Byung-Jun Yoon, Xihaier Luo, Vanessa López-Marrero, Patrick Johnstone, Shinjae Yoo, and Francis J. Alexander. Comparative performance evaluation of large language models for extracting molecular interactions and pathway knowledge, 2023. URL <https://arxiv.org/abs/2307.08813>.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library, 2019. URL <https://arxiv.org/abs/1912.01703>.
- David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluís-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. Carbon emissions and large neural network training, 2021. URL <https://arxiv.org/abs/2104.10350>.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners, 2019. URL <https://api.semanticscholar.org/CorpusID:160025533>.
- Vin Sachidananda, Jason S. Kessler, and Yi an Lai. Efficient domain adaptation of language models via adaptive tokenization, 2021. URL <https://arxiv.org/abs/2109.07460>.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL <https://aclanthology.org/P16-1162>.
- Jaime Sevilla, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos. Compute trends across three eras of machine learning, 2022. URL <https://arxiv.org/abs/2202.05924>.
- Nigam H. Shah, David Entwistle, and Michael A. Pfeffer. Creation and adoption of large language models in medicine. *JAMA*, 330(9):866, 2023. doi: 10.1001/jama.2023.14217.
- Or Sharir, Barak Peleg, and Yoav Shoham. The cost of training NLP models: A concise overview, 2020. URL <https://arxiv.org/abs/2004.08900>.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023a. doi: 10.1038/s41586-023-06291-2.
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaeckermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Towards expert-level medical question answering with large language models, 2023b. URL <https://arxiv.org/abs/2305.09617>.
- Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxin Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Øyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: an open corpus of three trillion tokens for language model pretraining research, 2024. arXiv preprint arXiv:2402.00159.

- Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science, 2022. URL <https://arxiv.org/abs/2211.09085>.
- Arun James Thirunavukarasu, Darren Shu Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Ting. Large language models in medicine. *Nature Medicine*, 29(8):1930–1940, 2023. doi: 10.1038/s41591-023-02448-8.
- Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C. Comeau, Rezarta Islamaj, Aadit Kapoor, Xin Gao, and Zhiyong Lu. Opportunities and challenges for ChatGPT and large language models in biomedicine and health, 2023. URL <https://arxiv.org/abs/2306.10070>.
- Together. Releasing 3B and 7B RedPajama-INCITE family of models including base, instruction-tuned & chat models, May 2023a. URL <https://www.together.ai/blog/redpajama-models-v1>.
- Together. RedPajama: An open dataset for training large language models, October 2023b. URL <https://github.com/togethercomputer/RedPajama-Data>.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models, 2023. URL <https://arxiv.org/abs/2302.13971>.
- George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, and et al. An overview of the BioASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16(1), 2015. doi: 10.1186/s12859-015-0564-6.
- Tom van Sonsbeek, Mohammad Mahdi Derakhshani, Ivona Najdenkoska, Cees G. M. Snoek, and Marcel Worring. Open-ended medical visual question answering through prefix tuning of language models, 2023. URL <https://arxiv.org/abs/2303.05977>.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. URL <https://arxiv.org/abs/1706.03762>.
- Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. <https://github.com/kingoflolz/mesh-transformer-jax>, May 2021.
- Mengqian Wang, Manhua Wang, Fei Yu, Yue Yang, Jennifer Walker, and Javed Mostafa. A systematic review of automatic text summarization for biomedical literature and EHRs. *Journal of the American Medical Informatics Association*, 28(10):2287–2297, 2021. doi: 10.1093/jamia/ocab143.
- Shibo Wang and Pankaj Kanwar. Bfloat16: The secret to high performance on cloud TPUs. <https://cloud.google.com/blog/products/ai-machine-learning/bfloat16-the-secret-to-high-performance-on-cloud-tpus>, 2019.
- Zuoxi Yang. Biomedical information retrieval incorporating knowledge graph for explainable precision medicine. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’20, page 2486, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380164. doi: 10.1145/3397271.3401458. URL <https://doi.org/10.1145/3397271.3401458>.
- Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D Manning, Percy Liang, and Jure Leskovec. Deep bidirectional language-knowledge graph pretraining, 2022a. URL <https://arxiv.org/abs/2210.09338>.
- Michihiro Yasunaga, Jure Leskovec, and Percy Liang. LinkBERT: Pretraining language models with document links, 2022b. URL <https://arxiv.org/abs/2203.15827>.

- Hye Sun Yun, Iain J. Marshall, Thomas Trikalinos, and Byron C. Wallace. Appraising the potential uses and harms of LLMs for medical systematic reviews, 2023. URL <https://arxiv.org/abs/2305.11828>.
- Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. Benchmarking large language models for news summarization, 2023. URL <https://arxiv.org/abs/2301.13848>.
- Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D. Manning, and Curtis P. Langlotz. Learning to summarize radiology findings, 2018. URL <https://arxiv.org/abs/1809.04698>.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models, 2023. URL <https://arxiv.org/abs/2303.18223>.
- Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. When does pretraining help? *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, 2021. doi: 10.1145/3462757.3466088.
- Weipeng Zhou, Majid Afshar, Dmitriy Dligach, Yanjun Gao, and Timothy Miller. Improving the transferability of clinical note section classification models with BERT and large language model ensembles. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 125–130, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.clinicalnlp-1.16. URL <https://aclanthology.org/2023.clinicalnlp-1.16>.
- Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2020. URL <https://arxiv.org/abs/1909.08593>.

A Examples Of Questions

1 MedMCQA Example Questions

Question	BioMedLM Answer
All of the following statements are true regarding hypertrophy, except:	A) Occurs due to synthesis and assembly of additional intracellular components. B) There is an increase in the size of the cells. C) Cells capable of division respond to stress by hypertrophy and hyperplasia. D) There is an increase in the number of cells.
A 19 year old female presents with pain in the neck for 5 days. She is not able to wear tie for her job because of neck pain. HO fatigue and lethargy for 10 days. She had flu like symptoms 20 days ago which resolved spontaneously. BP 110/80 mmHg, Pulse 48/min. Extremities are cold and dry. Neck is very tender. ECG normal. TSH is elevated. ESR 30 mm/hr. Next appropriate step	A) Atropine injection B) Levothyroxine administration C) Aspirin D) Increase iodine intake in food

Table A.1: Example Questions for MedMCQA

2 MedQA Example Question

Question	BioMedLM Answer
A 45-year-old woman presents to the emergency department with acute onset of severe right upper quadrant abdominal pain that radiates to the infrascapular region. Her medical history is significant for obesity, hypertension, obstructive sleep apnea, and gastric bypass surgery 2 years ago after which she lost 79 kg (150 lb). The patient complains of nausea and vomiting that accompanies the pain. Her temperature is 38.9°C (101.2°F), blood pressure is 144/88 mm Hg, heart rate is 76/min, and respiratory rate is 14/min (fever). Abdominal examination is significant for right upper quadrant tenderness along with guarding and cessation of inspired breath on deep palpation of the right upper quadrant. Which test should be ordered first for this patient?	A) Abdominal ultrasound B) CT scan of the abdomen C) MRI of the abdomen D) X-ray film of the abdomen

Table A.2: Example Question for MedQA

3 MMLU Example Question

Question	BioMedLM Answer
Zinc finger proteins and helix-turn-helix proteins are	A) types of DNA-binding proteins B) involved in the control of translation C) components of ribosomes D) part of the hemoglobin in blood cells

Table A.3: Example Question for MMLU

4 PubMedQA Example Question

Context:

To evaluate the degree to which histologic chorioamnionitis, a frequent finding in placentas submitted for histopathologic evaluation, correlates with clinical indicators of infection in the mother. A retrospective review was performed on 52 cases with a histologic diagnosis of acute chorioamnionitis from 2,051 deliveries at University Hospital, Newark, from January 2003 to July 2003. Third-trimester placentas without histologic chorioamnionitis ($n = 52$) served as controls. Cases and controls were selected sequentially. Maternal medical records were reviewed for indicators of maternal infection. Histologic chorioamnionitis was significantly associated with the usage of antibiotics ($p = 0.0095$) and a higher mean white blood cell count ($p = 0.018$). The presence of 1 or more clinical indicators was significantly associated with the presence of histologic chorioamnionitis ($p = 0.019$).

Question:

Does histologic chorioamnionitis correspond to clinical chorioamnionitis?

Yes/No/Maybe

Table A.4: Example Question for PubMedQA

5 BioASQ Example Question

Context:

NT-pro-BNP was significantly elevated postexercise in both adults and adolescents and remained above baseline at 24 h in both groups. NT-pro-BNP concentrations increased significantly (28 ± 17.1 vs 795 ± 823 ng x L, $P < 0.05$), whereas postrace cTnT were elevated in just five athletes (20%). [NT-pro-BNP] was observed immediately after the marathon (median [NT-pro-BNP] before: 39.6 pg ml⁻¹, after: 138.6 pg ml⁻¹, $p=0.003$) with a further increase on day one. [BNP] did not increase immediately after the marathon but increased on day one (median [BNP] before: 15 pg ml⁻¹, day one: 27.35 pg ml⁻¹, $p=0.006$). Pro-BNP was significantly increased immediately post-race (27 ± 21 vs 7 ± 2 pmol/L pre-race, $P < 0.007$), which 12-24 h later, decreased to 19 ± 14 pmol/L ($P = 0.07$ vs pre-race). The relatively high NT-proBNP levels after active recovery when psychophysical stress is higher, because of cycling and cold water immersion, suggest that not only endurance exercise, but also strenuous, stressful short exercise can induce an increase in NT-proBNP concentrations. Running a marathon significantly increases NT-pro-BNP levels in healthy adults. This increase could be partially attributed to cardiac stress. Increases in NT-proBNP can be found in a major part of obviously healthy athletes after prolonged strenuous exercise. The release of BNP during and after exercise may not result from myocardial damage but may have cytoprotective and growth-regulating effects. The different nature of exercise-induced increases in BNP and cardiac troponins has to be elucidated in the future. In healthy cyclists, transient increases in NT-pro-BNP and cTnT are more likely to reflect cardiac fatigue than injury. The rise in BNP in older athletes may reflect a reversible, mainly diastolic left ventricular dysfunction. Plasma BNP concentrations were higher in both the judo and marathon groups than in controls, and positively correlated with LV mass as well as with deceleration time. Such exercise significantly increased ANP and BNP levels in healthy men, and the increases could be partially attributed to myocardial damage during the race.

Question:

Does BNP increase after intensive exercise in athletes?

Yes/No

Table A.5: Example Question for BioASQ