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

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu
*Microsoft Research, October - 2022*

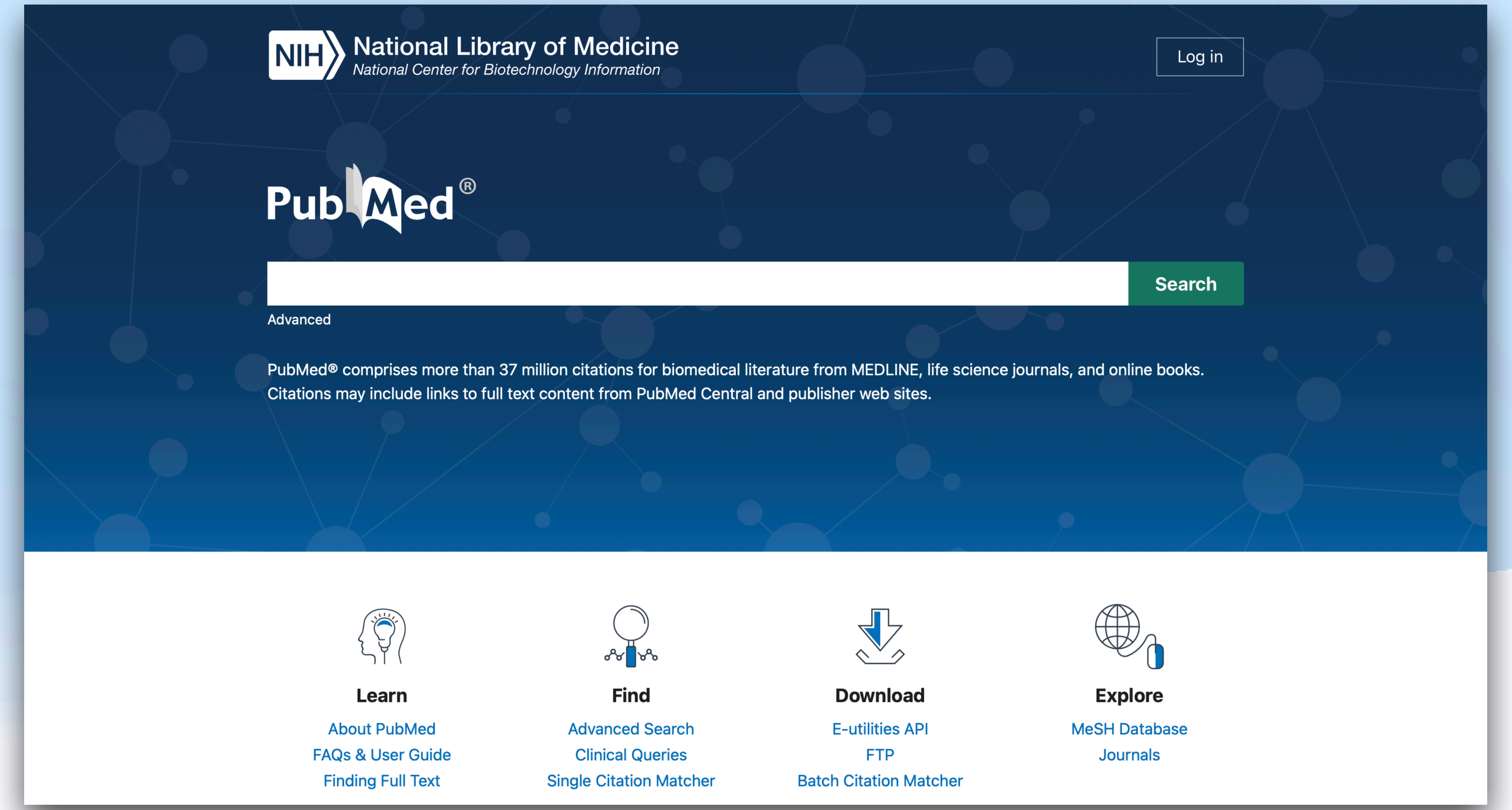
**Bhautik Lukhi**
**Advanced AI in Biomedicine(Graded)**
**21st November, 2024**

# Agenda

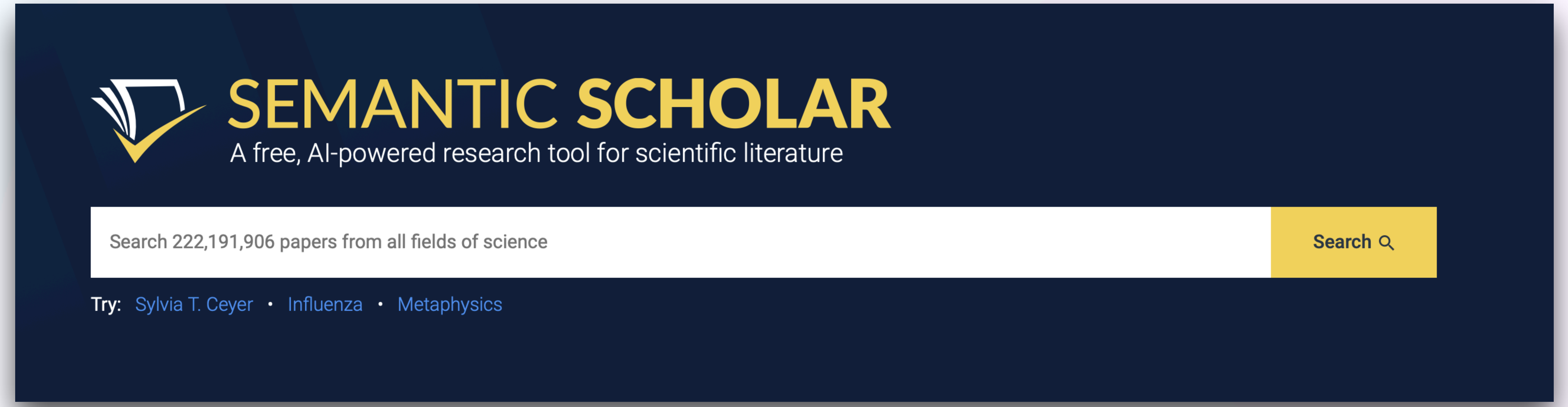
- Motivation
- Biomedical NLP Tasks
- BERT and GPT
- Introduction to BioGPT
- Architecture, Training and Fine-tuning
- Results on Biomedical Tasks
- BioGPT in Action

# Why

- Millions of Research Articles
- PubMed
- Semantic Scholar
- PMC
- Arxiv
- …..




PubMed



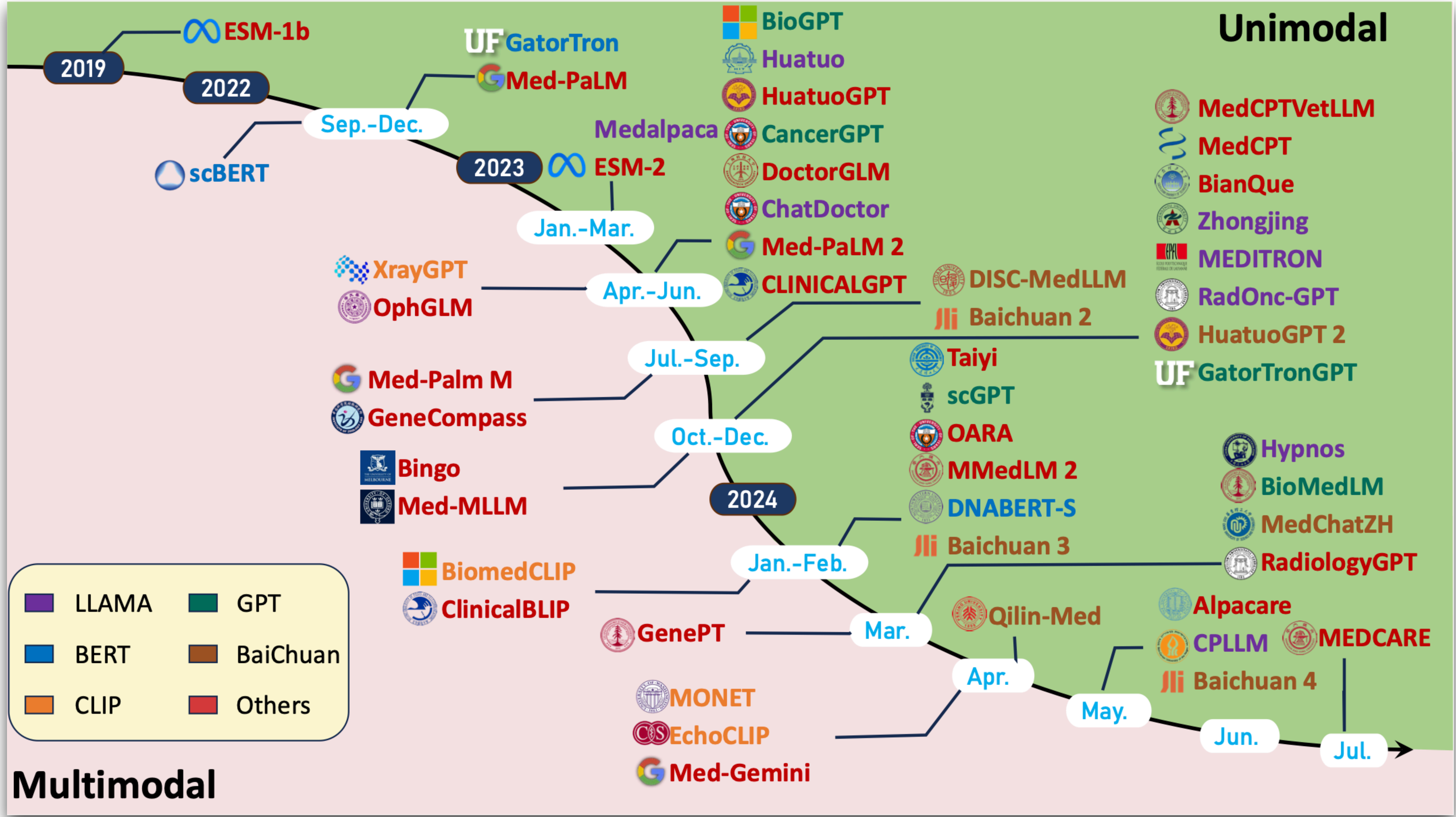

SEMANTIC SCHOLAR

# NLP in Biomedicine

- Great at extracting insights from structured/unstructured biomedical text.
- Exceptional capabilities in complex language understanding and generation tasks.
- **Applications:**
  - Drug discovery
  - Clinical therapy enhancement
  - Pathology research

# Biomedical NLP Tasks

- **Relation Extraction:** Identifying relationships between entities in text.
- **Question Answering:** Providing answers based on biomedical literature.
- **Document Classification**: Categorizing documents into predefined classes.
- **Text Generation:** Generating relevant biomedical text based on prompts.
- **Named Entity Recognition:** Identifying and classifying key entities.
- **Text Summarization:** Condensing lengthy articles into concise summaries.




A Survey for Language Models in Biomedicine (Wang et al. 2024)

# Pre-trained models for Bio-medicine

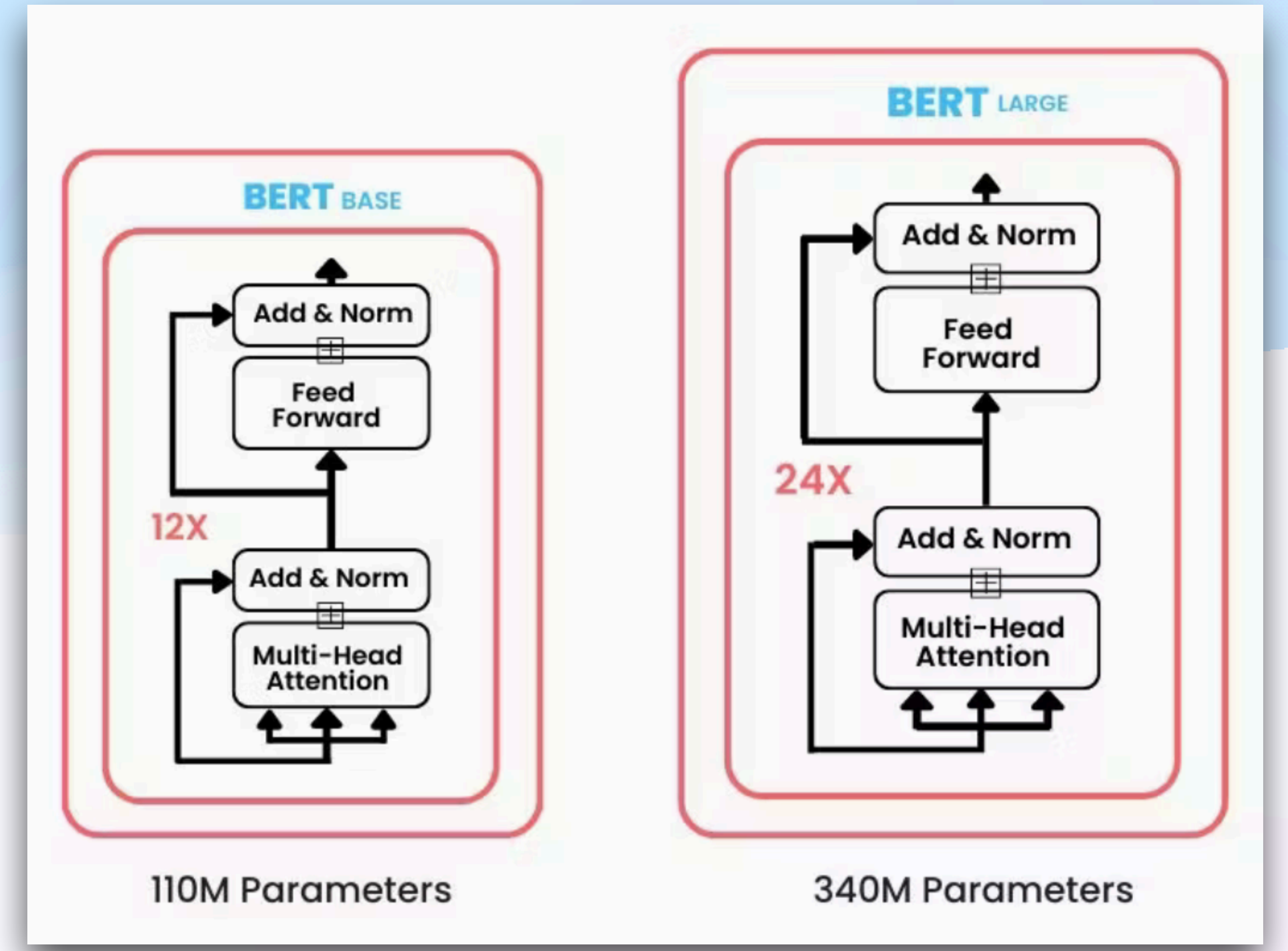
| Model | Pros | Cons |
|---|---|---|
| **BERT** | Pretrained on massive data | General Domain |
| **BioBERT** | Continue pre-trained on bio domain | Shared vocab with general domain |
| **BlueBERT** | Continue pre-trained on bio domain | Shared vocab with general domain |
| **SciBERT** | Pretrained on Science domain | Out-domain knowledge |
| **PubMedBERT** | Pretrained on bio domain | Encoder only architecture |
| **ELECTRAMed** | Pretrained on bio domain | Encoder only architecture |

*until November, 2022

# BERT and GPT

# BERT and GPT

- **BERT**: Pre-trained on English Wikipedia and BooksCorpus.
- **GPT**: Pre-trained Transformer pre-trained on BookCorpus.
- **Self-Supervision:**
  - Masked Language Modeling: Predicting masked words based on **full** context.
  - Causal language modeling: Predicting the next word of the sentence based **only** on the past.




BERT (Base) and BERT (Large).

# Where BERT Lacks?

- **BERT's Constraints:**
  - Better at understanding rather than generating text.
- **GPT's Generative Edge:**
  - Better at language generation through a causal language modeling.
  - GPT-2 and GPT-3 enhance performance on multiple downstream tasks.
  - Multi-task and also Few-shot learner.

# Why not use GPT Directly in Biomedicine?

- Even GPT-3 struggles with biomedical tasks due to the **Domain Shift.**
- **Previous Adaption Attempt:**
  - **DARE:** Pre-trained on limited data (**0.5M** abstracts) for data augmentation in relation extraction.
  - **But** results were limited.
  - **Task-Specific Adaptation(InstructGPT):** GPT model adapted for unconventional downstream clinical tasks.

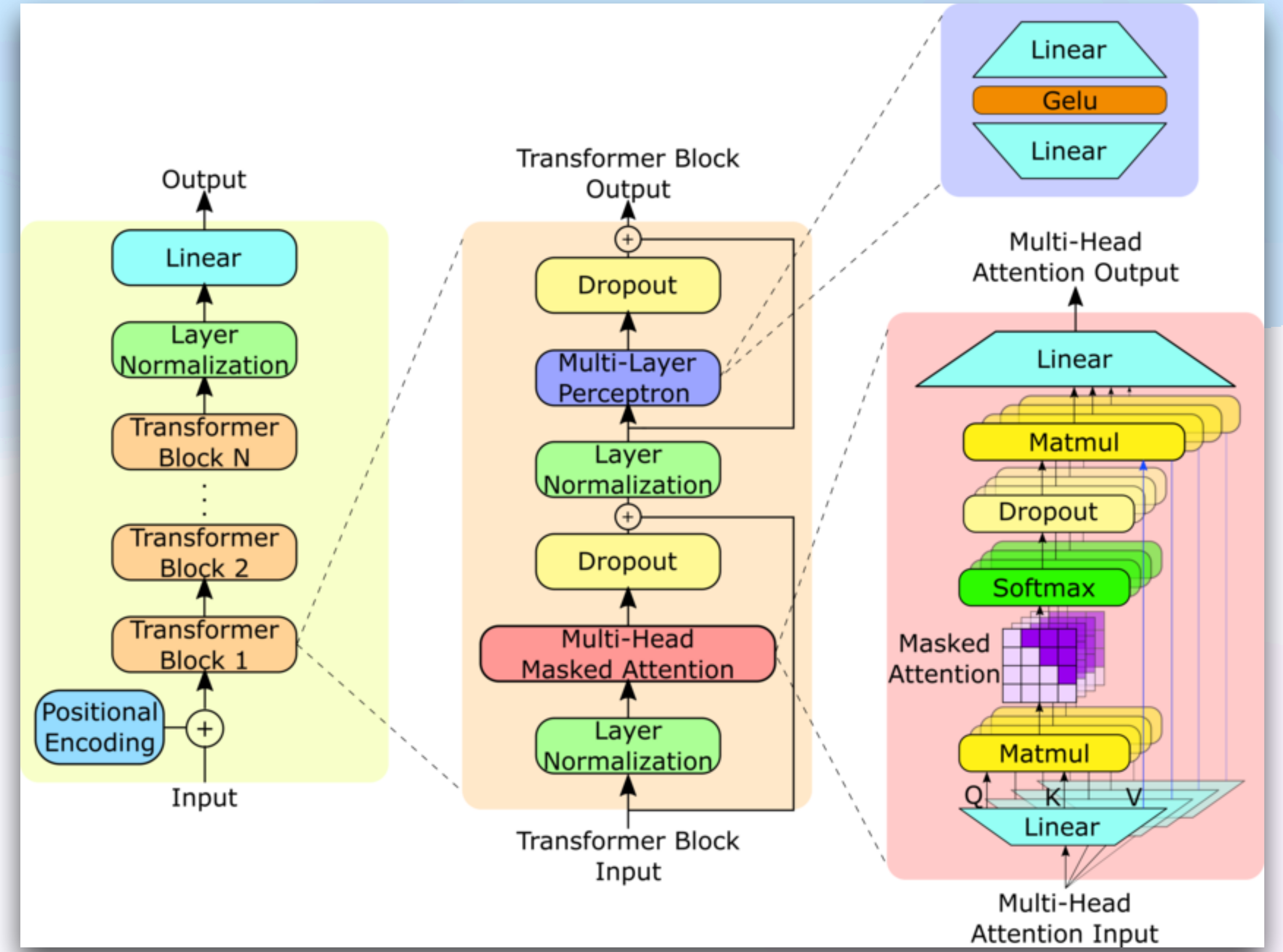
# Introduction to BioGPT

# What is BioGPT?

- A generative pre-trained Transformer for biomedical text, trained on **15M PubMed** abstracts.
- **Key Features:**
  - Combines generative modeling with biomedical relevance.
  - Can be used for Relation extraction, Question Answering, Document classification, and Text generation.

# Architecture

- Based on the GPT-2medium architecture.
- Features a Transformer decoder structure.
- **Key Components:**
  - Multi-head attention mechanism.
  - Trained end-to-end for optimized performance across various tasks.




GPT-2 Architecture (Yang et al., 2023)

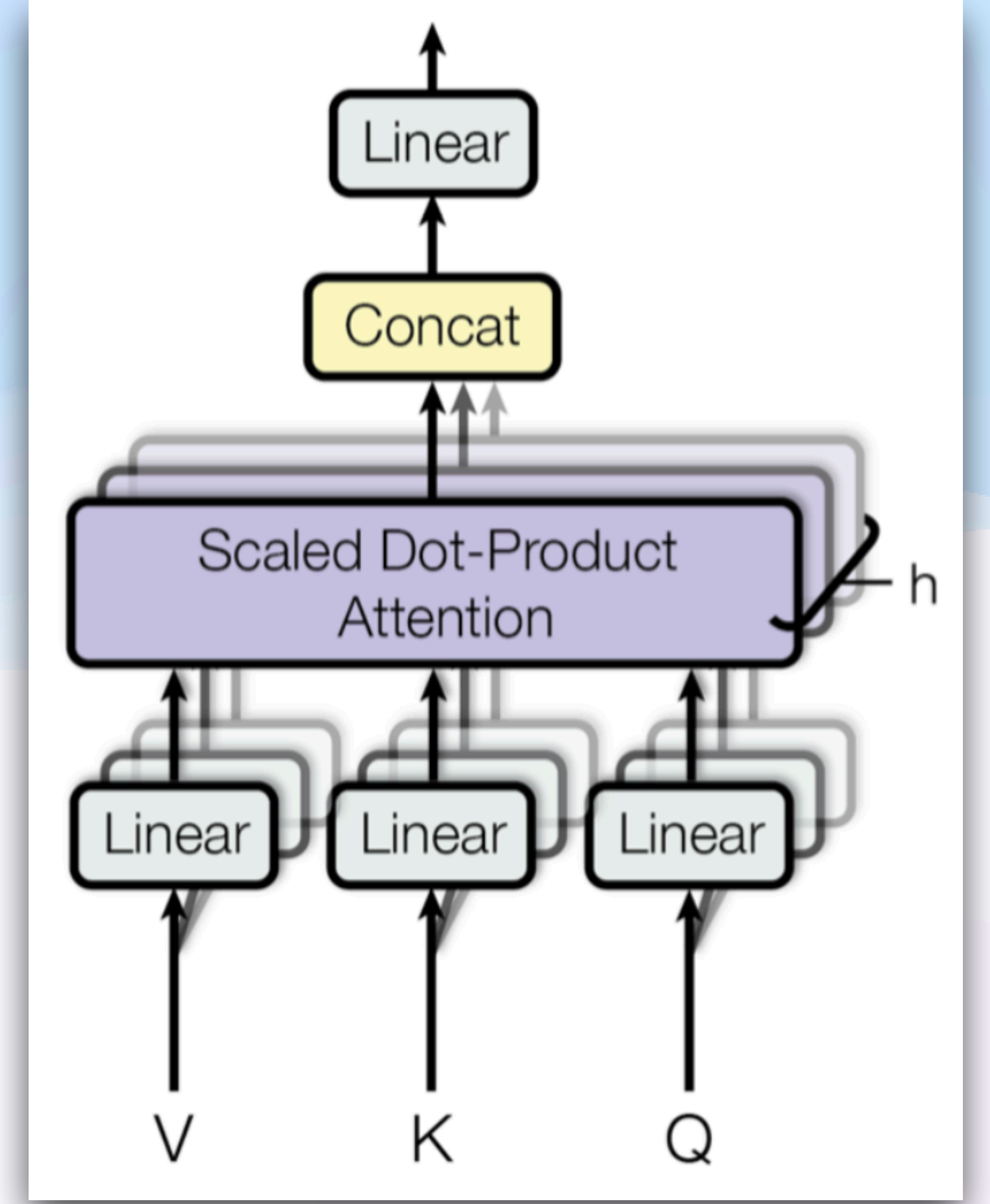
# Multi-Head Attention Mechanism

- Runs multiple attention mechanisms in parallel.
- Outputs are concatenated and linearly transformed to match dimensions.

$$Multihead(Q, K, V) = Concat(head_1, head_2, \ldots, head_h)W,$$

$$head_i = softmax\left(\frac{Q_i K_i^\top}{\sqrt{d}}\right)V_i$$

- $Q, K, V$: Queries, Keys, Values.
- $W$: Learnable parameter matrices.
- $d$: Dimensionality scaling factor.




Multi-Head Attention (Vaswani et al., 2017)

# Training and Fine-tuning

# Training Dataset

- Pre-trained on **15M PubMed** abstracts.
- Enables training effective language models with domain-specific knowledge.
- Bridge between general language understanding and biomedical text generation.
- **Key Characteristics:**
  - Diverse range of biomedical topics.
  - Continuously updated with newly published research.

# Training

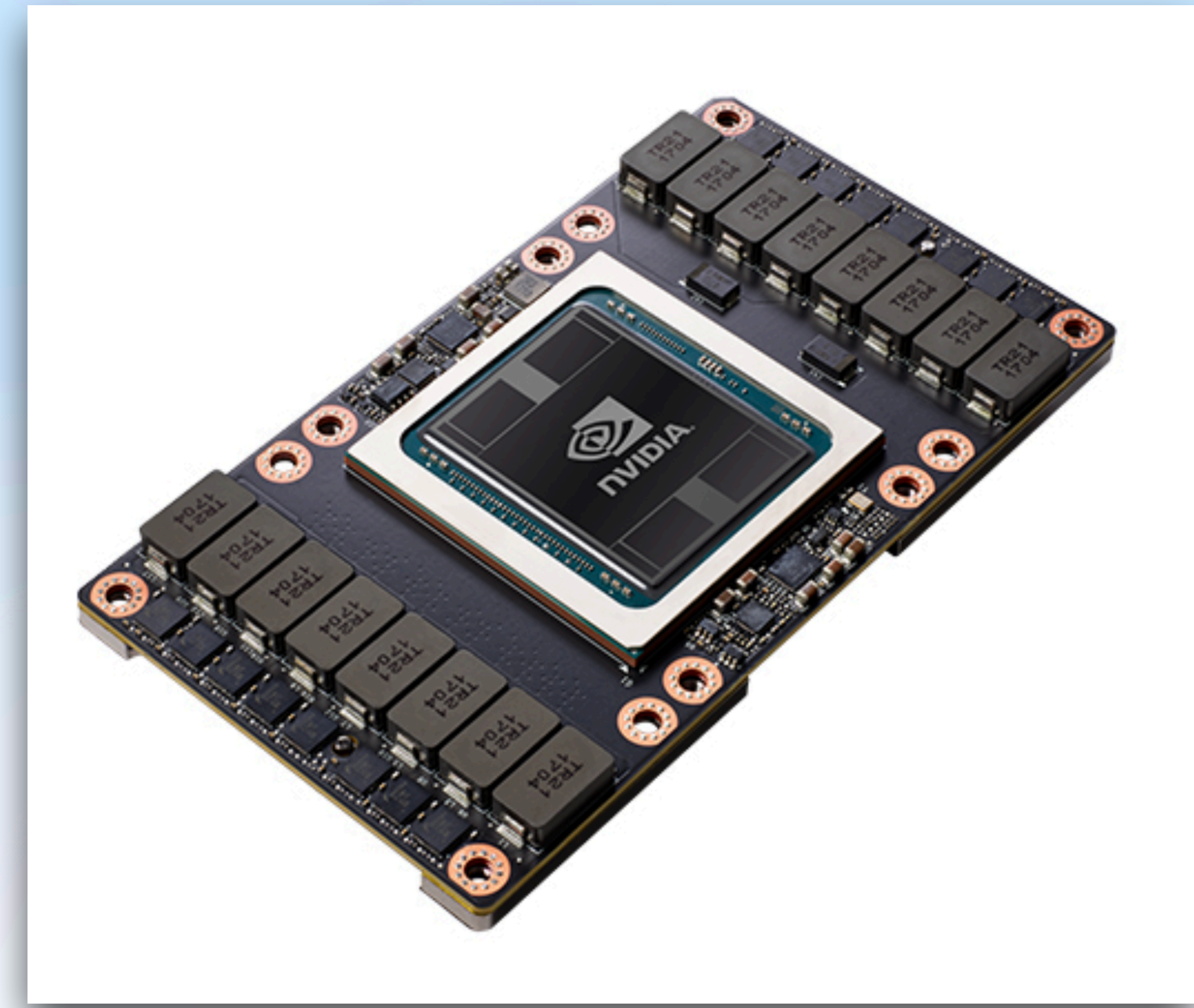
- Pre-trained using standard language modeling tasks.
- **Criteria:**
  - Minimize negative log-likelihood.

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

  - $D$: Dataset of sequences and $s_j$ : token
- Effective batch size of **524,288 tokens.**
- **Adam** optimizer with a learning rate schedule**.**
- Employed a **warm-up phase(20,000 steps)** to stabilize the training.

# Training Hardware

- Pre-trained on **8 NVIDIA V100** GPUs for 200,000 steps.



NVIDIA V100 FOR NVLINK, from NVIDIA website

# Vocabulary Development

- **Why Domain-Specific Vocabulary?**
  - Language model’s performance hinges on its vocabulary quality.
  - General vocabularies can complicate specialized biomedical terms.
- **Vocabulary Creation Process**
  - **Byte Pair Encoding (BPE):** Derives vocabulary directly from biomedical datasets.
  - **Final Vocabulary Size:** 42,384 tokens (50,257 for GPT)
- **Advantages:**
  - Better understanding of biomedical terminology.
  - More precise and contextually relevant text generation.

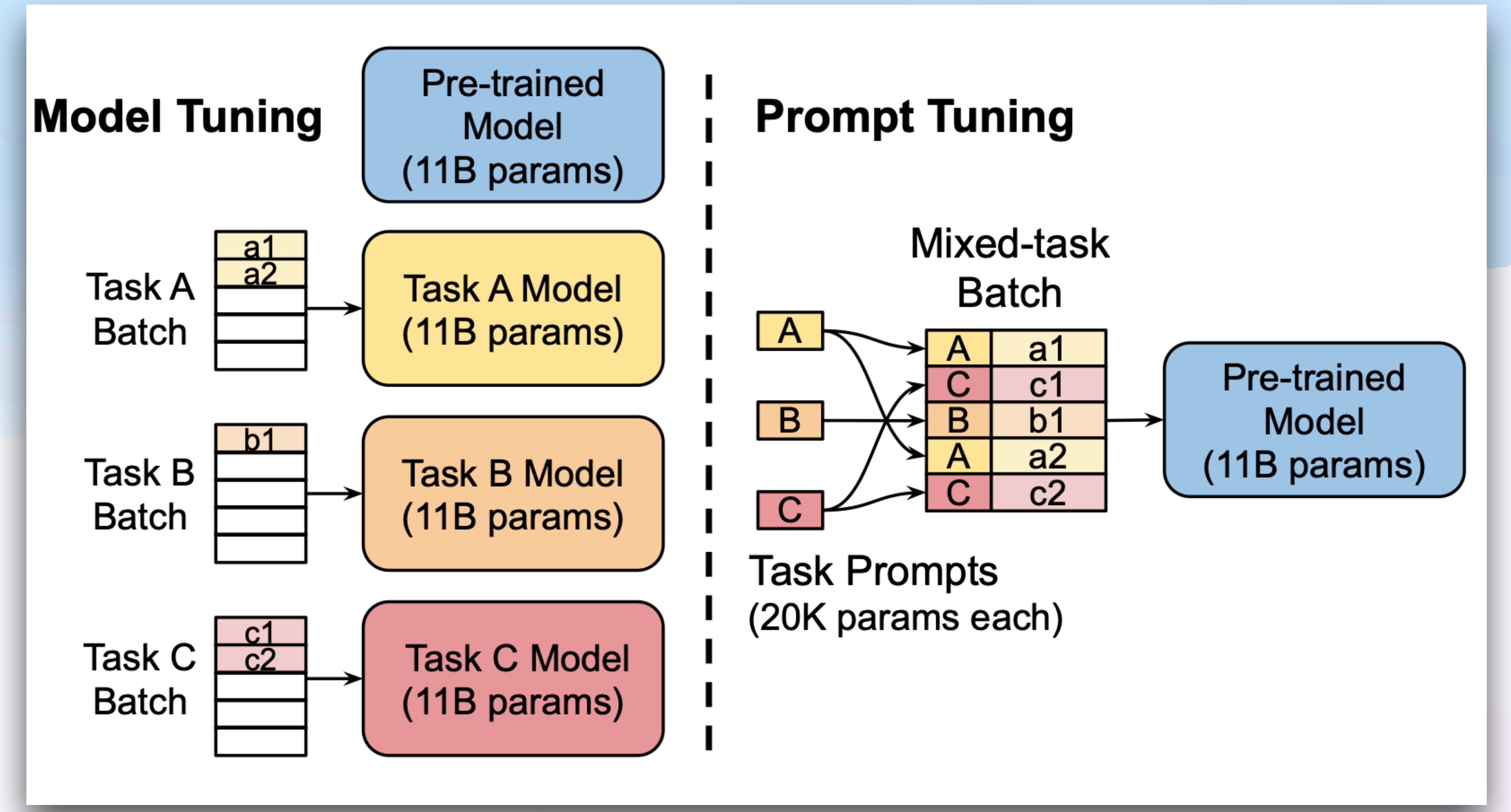
# Fine-tuning Method Overview

- To adapt BioGPT for downstream tasks.

- **Key Adaptation:**
  - Convert labels into natural language sequences.
  - Maintains consistency with pre-training task format.
  - Avoids structured formats or special tokens for smoother semantics.

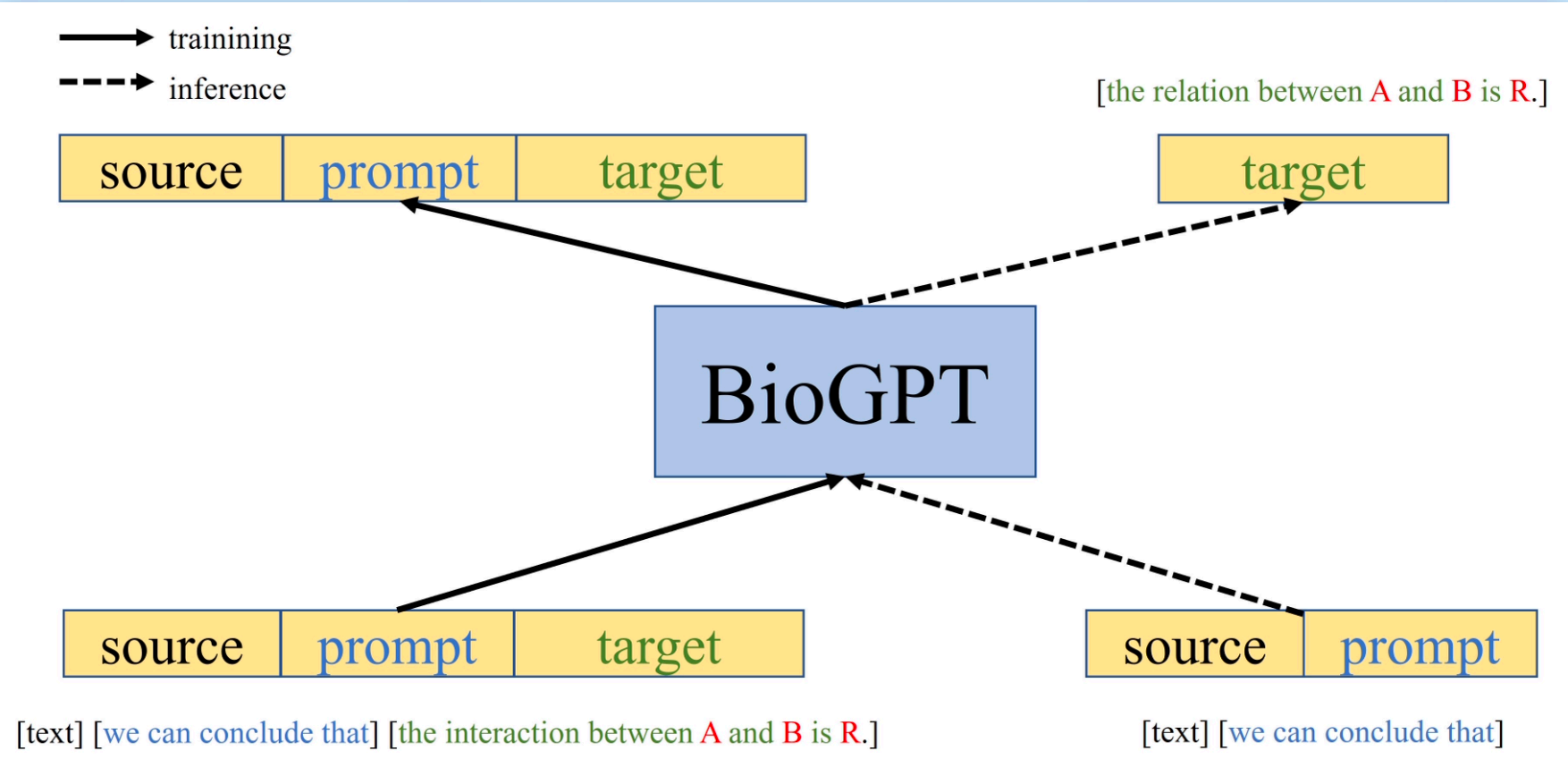
# Prompt Based Fine-tuning Framework

- Simply concatenating source and target is ineffective.
- **Workaround:** prompts to guide the model in generating task-specific output.
- **Soft Prompts:** Continuous embeddings (virtual tokens) learned end-to-end.
    - Virtual tokens are placed between source and target sequences.




Model Tuning vs Prompt Tuning (Lester et al., 2021)

# Prompt Based Fine-tuning Framework




Source: Paper

# Downstream Tasks

- **Relation Extraction:**
  - Key for biomedicine.
  - Focuses on generating relational triplets directly from text without intermediate annotations.
- **Question Answering (QA):**
  - Answering questions based on context.
  - Produce answer sequences, improving upon span prediction methods.
- **Document Classification:**
  - Classifying documents into predefined categories.
  - Leverages large pre-trained model for enhanced understanding and prediction.

# Downstream Tasks

| Task | Method | Dataset |
|---|---|---|
| **Relation Extraction** | GLRE, REBEL, seq2rel | KD-DTI, BC5CDR, DDI |
| **Question Answering** | QA-Net, LUKE, BERT, PubMedBERT, BioELECTRA, LinkBERT | PubMedQA, BioASQ |
| **Document Classification** | BERT, BlueBERT, SciBERT, SPECTER, PubMedBERT, BioELECTRA, LinkBERT | HoC, SciDocs |

Summary of the downstream tasks for evaluation

# BioGPT Variants

**BioGPT**

- 24-layer Transformer
- 347M parameters
- 15M PubMed abstracts
- Approx. 4B tokens

**$\text{BioGPT}_{\text{large}}$**

- 48- layer Transformer
- 1.57B parameters
- 15M PubMed abstracts + 6M PMC full paper
- Approx. 8B tokens

# Results

# 1. Relation Extraction Task

- Extracting relationships between entities(triplets) in a single pass, without needing intermediate steps.

| Model | Precision | Recall | F1 |
|---|---|---|---|
| GLRE(gt+pred) | 34.82 | 18.29 | 23.99 |
| GLRE(pred+pred) | 23.00 | 4.88 | 8.05 |
| GPT-2 | 43.92 | 32.55 | 37.39 |
| REBEL | 34.28 | 39.49 | 36.70 |
| $\text{REBEL}_{pt}$ | 40.94 | 21.20 | 27.94 |
| Seq2rel | 43.5 | 37.5 | 40.2 |
| **BioGPT** | 49.44 | 41.28 | **44.98** |
| **$\text{BioGPT}^{+}$** | 49.52 | 43.25 | **46.17** |

Results on **BC5CDR** chemical-disease-relation task

| Model | Precision | Recall | F1 |
|---|---|---|---|
| Transformer + PubMedBERT-attn | 25.35 | 24.14 | 24.19 |
| $\text{GPT-2}_{medium}$ | 30.53 | 27.87 | 28.45 |
| REBEL | 32.36 | 29.58 | 30.39 |
| $\text{REBEL}_{pt}$ | 35.73 | 32.61 | 33.32 |
| **BioGPT** | 40.00 | 39.72 | **38.42** |

Results on **KD-DTI** drug-target-interaction task

| Model | Precision | Recall | F1 |
|---|---|---|---|
| $\text{GPT-2}_{medium}$ | 23.39 | 31.93 | 24.68 |
| REBEL | 35.36 | 28.64 | 28.27 |
| $\text{REBEL}_{pt}$ | 46.59 | 39.60 | 40.56 |
| **BioGPT** | 41.70 | 44.75 | **40.76** |

Results on **DDI** drug-drug-interaction task

# 2. Question Answering

- **Goal:** Answer questions using reference context.
- **Labels:** Yes, No, Maybe.

| Model | Accuracy |
|---|---|
| PubMedBERT | 55.8 |
| BioELECTRa | 64.2 |
| $\text{BioLinkBERT}_{\text{base}}$ | 70.2 |
| $\text{BioLinkBERT}_{\text{large}}$ | 72.2 |
| **BioGPT** | **78.2** |

Results on **PubMedQA** question answering task

# 3. Document Classification

- **Goal:** Classify document type based on text.
- **Target Sequence Format:** The type of this document is 'label'.

| Model | F1 |
|---|---|
| BioBERT | 81.54 |
| PubMedBERT | 82.32 |
| $\text{PubMedBERT}_{\text{large}}$ | 82.70 |
| $\text{BioLinkBERT}_{\text{base}}$ | 84.35 |
| $\text{GPT-2}_{\text{medium}}$ | 81.84 |
| **BioGPT** | **85.12** |

Results on **HoC** document classification task

# Results Summary

**Relation Extraction:**

- Drug-Target Interaction (KD-DTI)
- Chemical-Disease Interaction (BC5CDR)
- Drug-Drug Interaction (DDI)
- **Up to 4%** improvement over all methods

**Question Answering:**

- PubMedQA
- **6.0%** improvement over previous best

**Document Classification:**

- HoC
- **3.28%** improvement over previous ones

# BioGPT in Action

# Drug-Target Interaction Extraction

The Janus family kinases (Jaks), Jak1, Jak2, Jak3, and Tyk2, form one subgroup of the non-receptor protein tyrosine kinases. They are involved in cell growth, survival, development, and differentiation of a variety of cells but are critically important for immune cells and hematopoietic cells. Data from experimental mice and clinical observations have unraveled multiple signaling events mediated by Jak in innate and adaptive immunity. Deficiency of Jak3 or Tyk2 results in defined clinical disorders, which are also evident in mouse models. A striking phenotype associated with inactivating Jak3 mutations is severe combined immunodeficiency syndrome, whereas mutation of Tyk2 results in another primary immunodeficiency termed autosomal recessive hyperimmunoglobulin E syndrome. In contrast, complete deletion of Jak1 or Jak2 in the mouse are not compatible with life and, unsurprisingly, do not have counterparts in human disease. However, activating mutations of each of the Jaks are found in association with malignant transformation, the most common being gain-of- function mutations of Jak2 in polycythemia vera and other myeloproliferative disorders. Our existing knowledge on Jak signaling pathways and fundamental work on their biochemical structure and intracellular interactions allow us to develop new strategies for controlling autoimmune diseases or malignancies by developing selective Jak inhibitors, which are now coming into clinical use. Despite the fact that Jaks were discovered only a little more than a decade ago, at the time of writing there are 20 clinical trials underway testing the safety and efficacy of Jak inhibitors.

BioGPT

- **Output:** the interaction between pnus156804 and janus kinase 3 (jak 3) is inhibitor.
- **Structured Output:** <pnus156804, janus kinase 3(jak 3), inhibator >

**Task**: Find <drug, target, interaction> triplet given the document

# Drug-Drug Interaction Extraction

An inhibitor of CYP2C8 (such as gemfibrozil) may increase the AUC of rosiglitazone and an inducer of CYP2C8 (such as rifampin) may decrease the AUC of rosiglitazone. Therefore, if an inhibitor or an inducer of CYP2C8 is started or stopped during treatment with rosiglitazone, changes in diabetes treatment may be needed based upon clinical response.

BioGPT →

- **Output:** the interaction between gemfibrozil and rosiglitazone is mechanism; the interaction between rifampin and rosiglitazone is mechanism.

- **Structured output:** (gemfibrozil, rosiglitazone, mechanism), (rifampin, rosiglitazone, mechanism)

Task : Find <drug, drug, interaction> triplet given the document

# Question Answering

**Question:** Do some US states have higher / lower injury mortality rates than others?

**Context:** this article examines the hypothesis that the six u.s. states with the highest rates of road traffic deaths (group 1 states) also had above-average rates of other forms of injury such as falling, poisoning, drowning, fire, suffocation, homicide, and suicide, and also for the retail trade and construction industries. the converse, second hypothesis, for the six states with the lowest rates of road traffic deaths (group 2 states) is also examined. data for these 12 states for the period 1983 to 1995 included nine categories of unintentional and four categories of intentional injury, seventy-four percent of the group 1 states conformed to the first hypothesis, and 85% of the group 2 states conformed to the second hypothesis, answer: group 1 states are likely to exhibit above-average rates for most other categories of injury death, whereas group 2 states are even more likely to exhibit below-average rates for most other categories of injury death.

**Ground truth:** Yes

**BioGPT:** the answer to the question given the context is **Yes.**

# Zero-shot QA

**Question:** Can we measure mesopic pupil size with the cobalt blue light slit-lamp biomicroscopy method?

**Context:** [tl;dr: Some background introduction] The aim of this work is to assess a previously described slit-lamp biomicroscopy-based method (SLBM) for measuring pupil diameter and compare it to Colvard infrared pupillometry (CIP). Two examiners performed three repeated measurements with each instrument in 40 healthy eyes. We determined the agreement of SLBM and CIP, intraobserver and interobserver repeatabilities, and interobserver concordance (kappa) and SLBM ability for detecting pupil sizes over 6.0 mm. The mean (±standard deviation [SD]) pupil diameter was 5.81 ± 0.70 mm with SLBM and 6.26 ± 0.68 mm with CIP (p=0.01) averaging both examiner's results. Mean differences between the SLBM and CIP were -0.60 mm and -0.30 mm for each examiner using the average of the three readings (p=0.02), and they were very similar using the first reading. Intraobserver reproducibility: the width of the 95% LoA ranged from 1.79 to 2.30 mm. The ICCs were 0.97 and 0.92 for SLBM, and 0.96 and 0.90 for CIP. Interobserver reproducibility, the width of the LOA ranged from 1.82 to 2.09 mm. Kappa statistics were 0.39 and 0.49 for the first and mean SLBM readings, respectively, and 0.45 for both the first and mean CIP readings. Sensitivity and specificity of SLBM for detection of pupils larger than 6 mm ranged from 55.56% to 73.68% and from 76.19% to 95.45%, respectively. The best tradeoff between sensitivity and specificity ranged from 5.4 mm to 6.2 mm.

**Ground truth:** No

**BioGPT:** The answer is that SLBM **is not a good alternative** to CIP for measuring pupil size.

# Text Generation
## Common Words

| Input | Model | Text |
| --- | --- | --- |
| Bicalutamide | GPT-2 | Bicalutamide, a synthetic bicalutamide derivative, was found to increase the cellular proliferation of C. elegans cells. This effect was not observed with the other synthetic bicalutamide derivatives. This is in line with our previous work which showed that bicalutamide increases the proliferation of C. elegans cells but not in the presence of synthetic bicalutamide. |
| | BioGPT | Bicalutamide (Casodex) is an androgen receptor (AR) antagonist approved for the treatment of metastatic castration-resistant prostate cancer (mCRPC) in patients who have progressed on or are ineligible for docetaxel chemotherapy, as well as for the treatment of early-stage prostate cancer in men who have not progressed on or are ineligible for docetaxel chemotherapy, as well as for the treatment of metastatic castration-sensitive prostate cancer (mCSPC) in men who have not progressed on or are ineligible for docetaxel chemotherapy. |

GPT-2 vs BioGPT when given relatively common names as input

# Text Generation

## Uncommon Words I

| | | |
|---|---|---|
| Xylazine | GPT-2 | Xylazine, the "active ingredient" in "bath salts" (see above) has been linked to numerous deaths, including those of former University of California, Berkeley, professor Andrew Anglin, who posted a video of himself having a bath filled with it. |
| | BioGPT | Xylazine is an alpha 2-adrenoceptor agonist that has been used as a sedative and analgesic in veterinary medicine for many years, but its effects on the cardiovascular system have not been extensively studied in the dog, and its effects on the central nervous system (CNS) have not been well characterized in the dog, despite the fact that xylazine has been widely used as a sedative and analgesic in veterinary medicine for more than 30 years. |
| Psoralen | GPT-2 | Psoralen, the city of the spiritualist, and had been given to him by God, who wished to show him the way of salvation, and to draw him back to the true life of Christ, and to bring him into communion with his fellow-man. |
| | BioGPT | Psoralen photochemotherapy (PUVA) is a well-established treatment for psoriasis, but its use is limited by the risk of skin cancer, particularly squamous cell carcinoma (SCC) of the head and neck (H & N), which is the most common site of skin cancer in the United States (US) |

GPT-2 vs BioGPT when given somewhat uncommon names as input

# Text Generation

## Uncommon Words II

| | | |
|---|---|---|
| CP-673451 | GPT-2 | CP-673451) and 4-5-6-3-1-2-3-4-6-7-8-9-10-11-12-13-14-15-16-17-18-19-20-21-22-23-24-25-26-27-28-29-30-31-32-33-34-35-36-37-38-39-40-41-42-43-44-45-46-47-48-49-50... |
| | BioGPT | CP-673451 is a potent, selective, and orally active inhibitor of human neutrophil elastase (HNE) and human cathepsin G (CatG) with in vitro and in vivo anti-inflammatory activity in a variety of animal models of inflammation and in a model of acute lung injury (ALI) in the rat induced by intratracheal instillation of lipopolysaccharide (LPS) and tumor necrosis factor-alpha (TNF-alpha), a model of acute lung injury (ALI) in which neutrophils play an important role. |
| BIIB-021 | GPT-2 | BIIB-021A-2-1.pdf. Wright, C. (2002). "The War on Terror: A Legal Theory of Terrorist Attacks," in L.E.R. Wright, S.R. Warrick, and J.H. Rothman (Eds.), The War on Terror: Essays in Honor of Michael T. Klare (New York: Oxford University Press), 143-57. |
| | BioGPT | BIIB-021 is a novel, orally active, non-peptide bradykinin B2 receptor antagonist with potent and long-lasting anti-inflammatory activity in animal models of acute and chronic inflammation and in a rat model of adjuvant-induced arthritis (AIA), an animal model of rheumatoid arthritis (RA), and in a rat model of collagen-induced arthritis (CIA), an animal model of collagen-induced arthritis (CIA), in which arthritis is induced by immunization with bovine type II collagen (CII). |

GPT-2 vs BioGPT when given somewhat uncommon names as input

# Text Generation

## Keywords of interest

| Input | Model | Text |
|---|---|---|
| COVID-19 | GPT-2 | COVID-19-0067-7 Hip-hop-romance-detective-franchise-unveiled-in-coco-houston http://www.pinterest.com/pin/71178278499149422/ Black Widow: World War II's Greatest Widow? http://www.pinterest.com/pin/71177842668650112/ |
| | BioGPT | COVID-19 is a global pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the causative agent of coronavirus disease 2019 (COVID-19), which has spread to more than 200 countries and territories, including the United States (US), Canada, Australia, New Zealand, the United Kingdom (UK), and the United States of America (USA), as of March 11, 2020, with more than 800,000 confirmed cases and more than 800,000 deaths. |
| SARS-CoV-2 | GPT-2 | SARS-CoV-2-CKM-T1-1/pdf/T1/SARS-CoV-2-CKM-T1-1.pdf Cancer http://www.cancer.gov/cancer topics/programs/tox/tox-programs.html Other Research http://www.cdc.gov/cancer/cancer/index.html |
| | BioGPT | SARS-CoV-2 is the causative agent of COVID-19, a severe acute respiratory syndrome (SARS) that has infected more than 390,000 people worldwide and killed more than 250,000 people. |

GPT-2 vs BioGPT when manually given keywords of interest (COVID-19 related terms)

# Scaling to Larger Size

## BioGPT$_{large}$

- Developed on **GPT-2 XL** architecture, **1.5B** parameters.
- Fine-tuned and evaluated for **enhanced performance** on downstream tasks.

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

BioGPT-Large fine-tuned on downstream tasks

# Conclusion

- Built on the GPT-2 backbone, pre-trained on **15M PubMed** abstracts.
- **Pioneer** to adapt GPT effectively in Biomedicine domain.
- **Outperforms** GPT-2 in biomedical text generation.
- **State-of-the-Art** results on:
  - 3 relation extraction tasks.
  - 1 question answering task.
- Larger-scale BioGPT model on expanded biomedical datasets.

# References


- https://arxiv.org/pdf/2210.10341
- https://pubmed.ncbi.nlm.nih.gov/
- https://www.semanticscholar.org/
- https://arxiv.org/pdf/2409.00133
- https://vitalflux.com/bert-vs-gpt-differences-real-life-examples/
- https://www.researchgate.net/figure/GPT-2-model-architecture-The-GPT-2-model-contains-N-Transformer-decoder-blocks-as-shown_fig1_373352176
- https://arxiv.org/pdf/1706.03762
- https://www.nvidia.com/de-de/data-center/

# Any Questions?

## Thank you for listening!