NLP SUMMIT

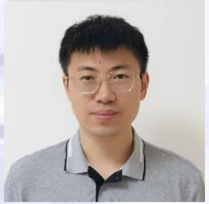
HEALTHCARE

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

Renqian Luo

Researcher

Microsoft Research



Yingce Xia

Principle Researcher

Microsoft Research

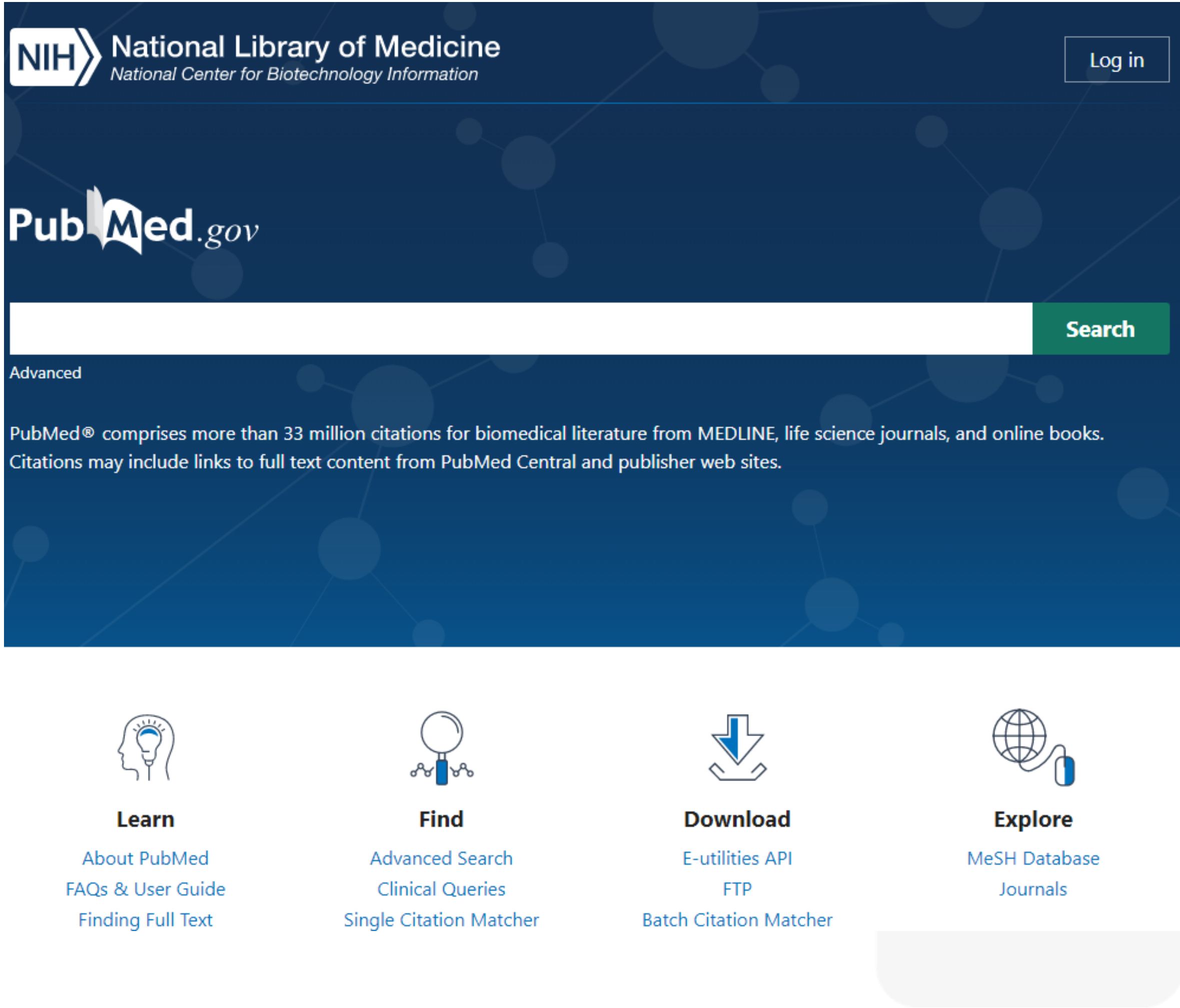
What we did:

A GPT model trained on biomedical literature to assist AI4Science research

Why

- Numerical documents

- PubMed
- PMC
- Semantic Scholar
- Arxiv
- …



NIH National Library of Medicine
National Center for Biotechnology Information

PubMed.gov

PubMed® comprises more than 33 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full text content from PubMed Central and publisher web sites.

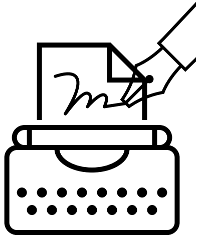
NLP SUMMIT

Pre-trained language model for biomedicine and life science

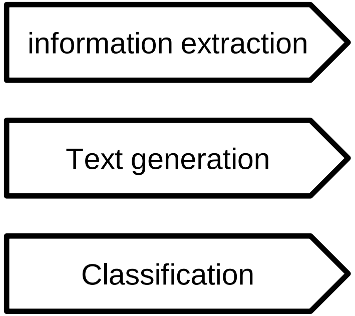
Model	Advantage	Disadvantage
BERT	Pretrained on massive data	General domain
BioBERT	Continue pretrained on bio domain	Shared vocab with general domain
BlueBERT	Continue pretrained on bio domain	Shared vocab with general domain
SciBERT	Pretrained on science domain	Out-domain knowledge
PubMedBERT	Pretrained on bio domain	Encoder only arch for understanding
ELECTRAMed	Pretrained on bio domain	Encoder only arch for understanding

Why GPT

Powerful
Generative ability



Multi-task learner

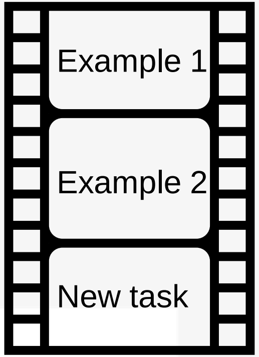


information extraction

Text generation

Classification

Few-shot learner



Example 1

Example 2

New task

Our model

BioGPT

24-layer Transformer #param=345M

30M PubMed abstracts (approx. 4B tokens)

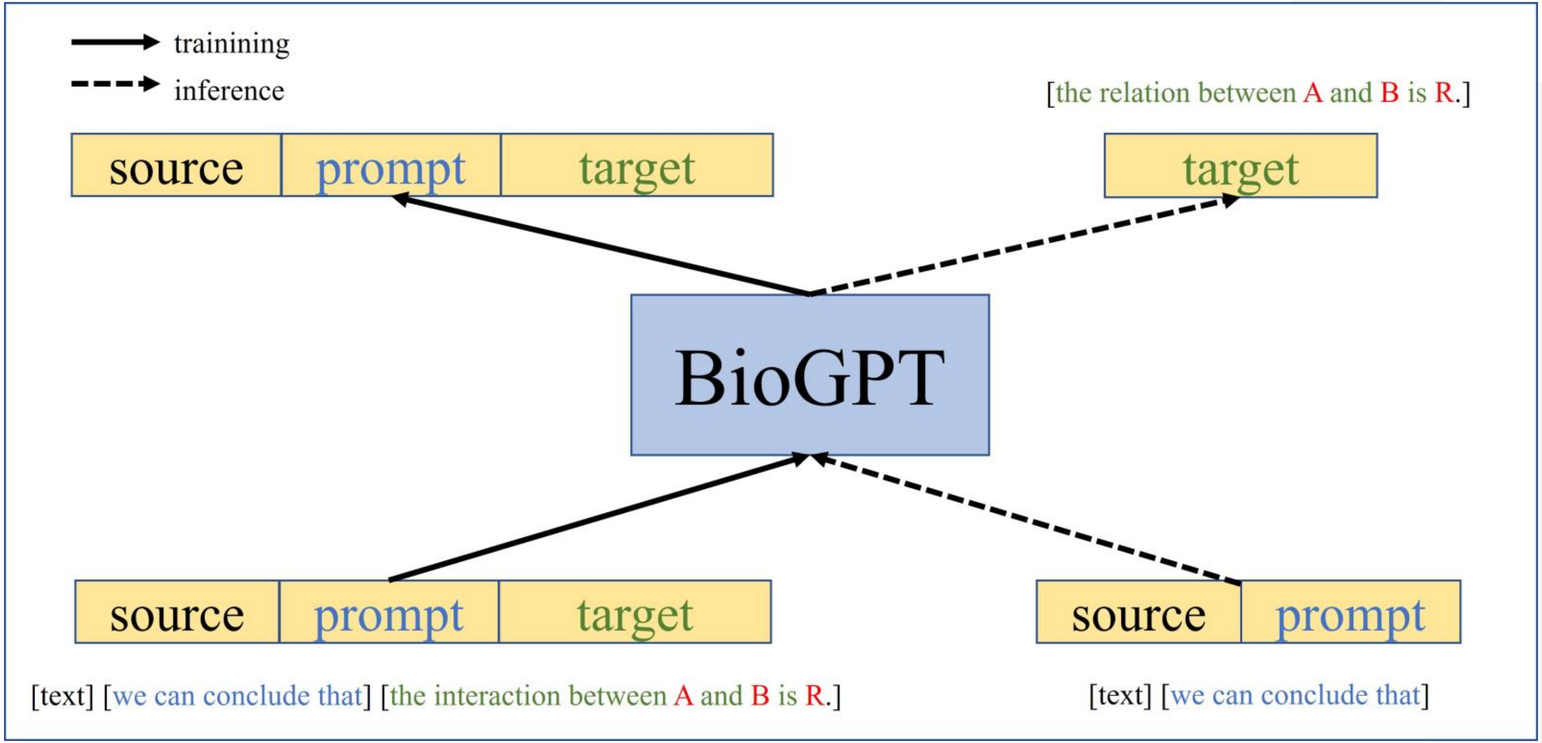
BioGPT-large

48-layer Transformer #param=1.57B

30M PubMed abstracts + 6M PMC full paper (approx. 8B tokens)

- Normalize numbers to digits
- Structuralize the literature
- <title> title </title>, <abstract> abstract </abstract>, <introduction> introduction </introduction>

Prompt Based Finetuning Framework


training
inference
[the relation between A and B is R.]
source
prompt
target
target
BioGPT
source
prompt
target
source
prompt
[text] [we can conclude that] [the interaction between A and B is R.]
[text] [we can conclude that]

NLP SUMMIT

Results

Relation Extraction

- Drug Target Interaction
- Chemical Disease Relation
- Drug Drug Interaction

- Better than all previous methods up to 4% improvements

Question Answering

- PubMedQA

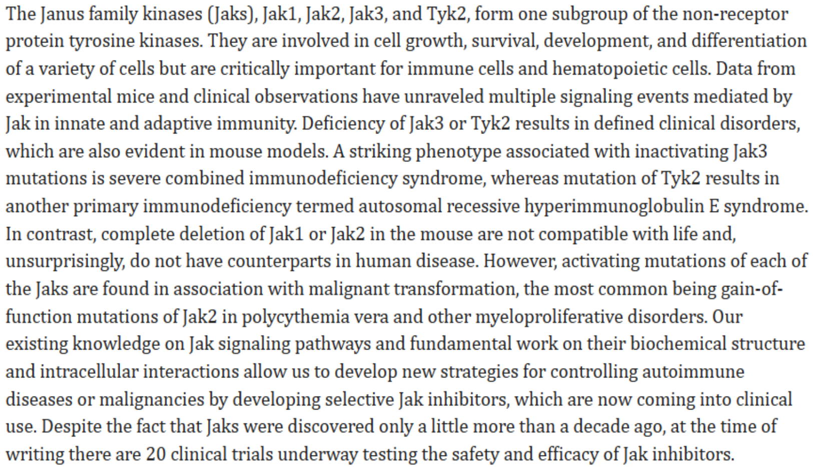
- Human parity

Document Classification

- HoC

- 3.28% improvement over previous model

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 pnu156804 and janus kinase 3 (jak-3) is inhibitor.
- *Structured output*: (pnu156804, janus kinase 3 (jak-3), inhibitor)

Task:
find the $\langle \text{drug}, \text{target}, \text{interaction} \rangle$ triplets in 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 the $\langle \text{drug}, \text{drug}, \text{interaction} \rangle$ triplets in the document

NLP SUMMIT

Question Answering

- Question: Do some u.s. 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.

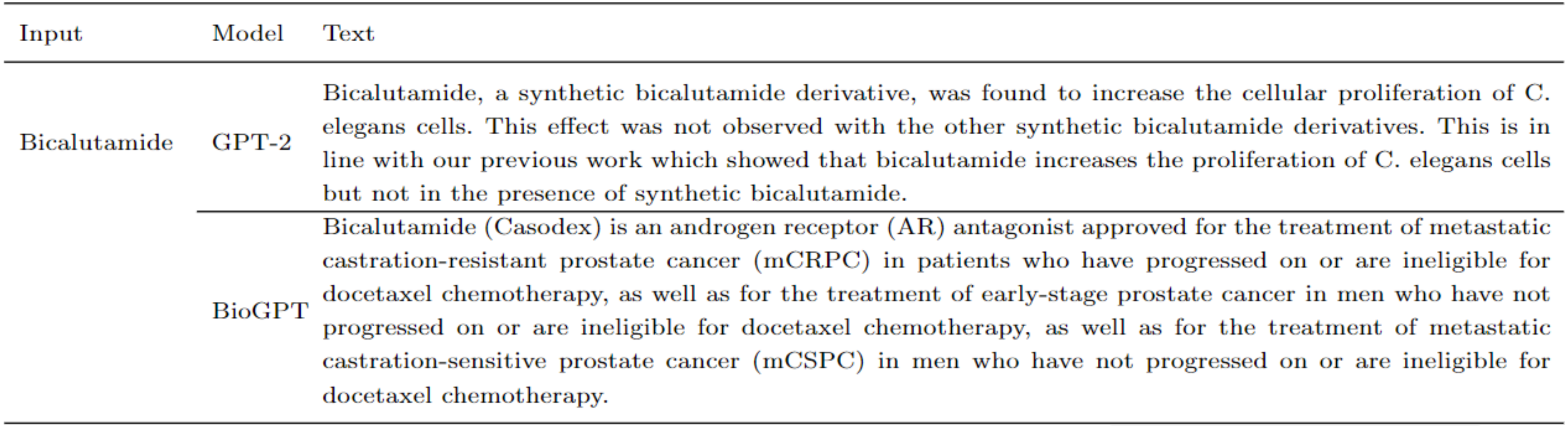
NLP SUMMIT

Zero-shot learning

- 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

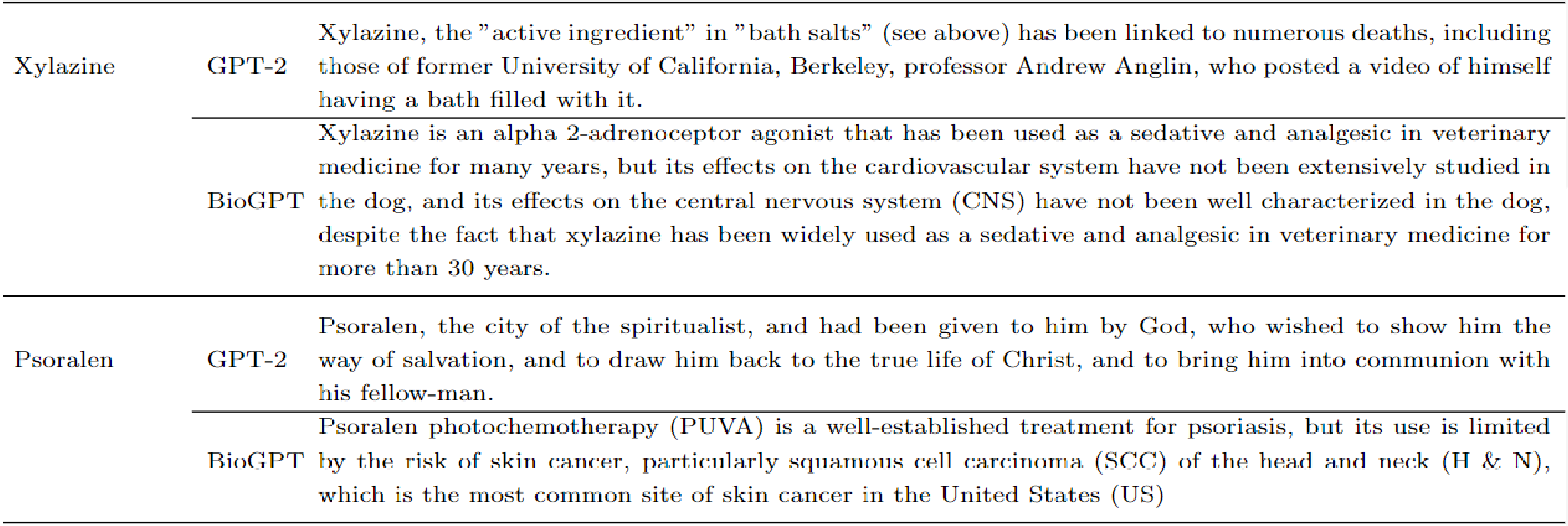
- Common words



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.

Text Generation

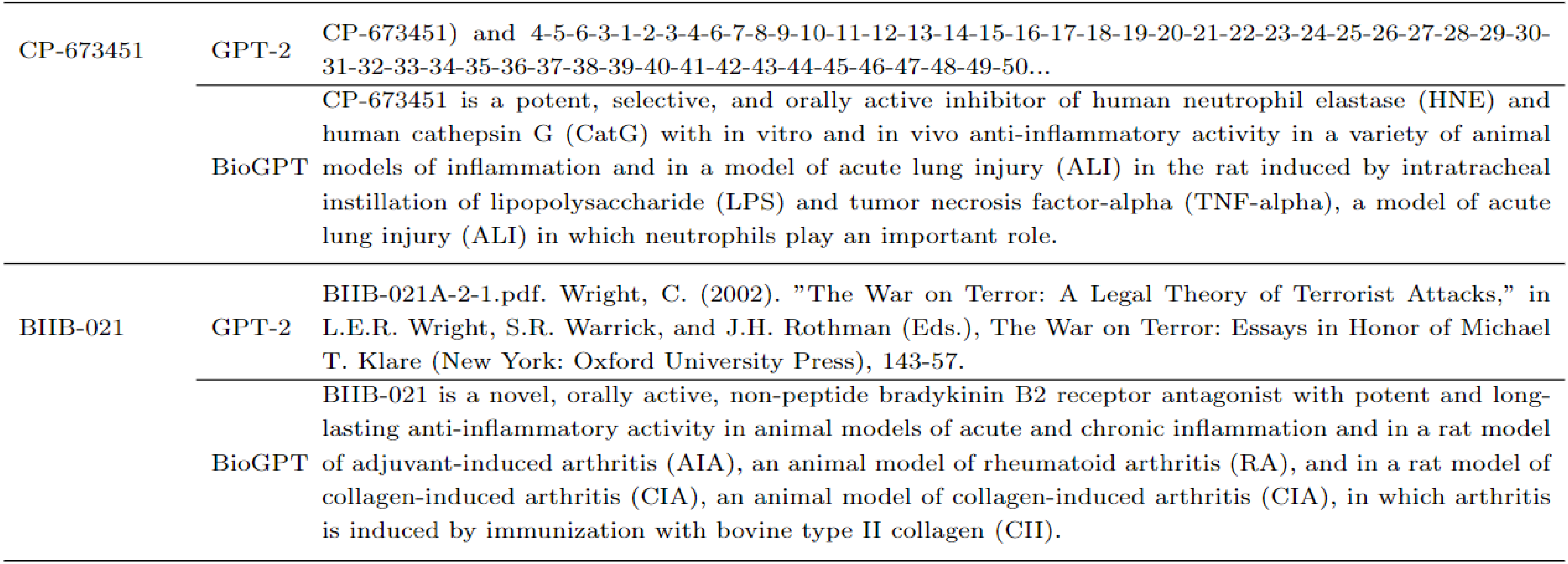
- Less common/Uncommon words



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)

Text Generation

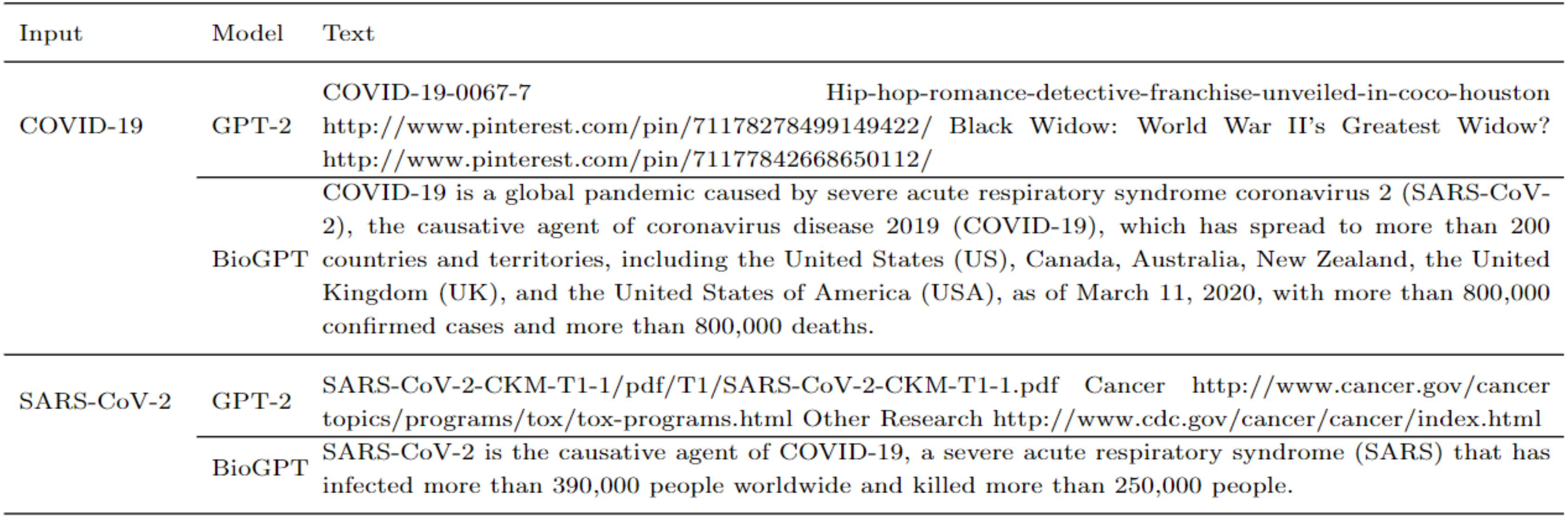
- Less common/Uncommon words



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).

Text Generation

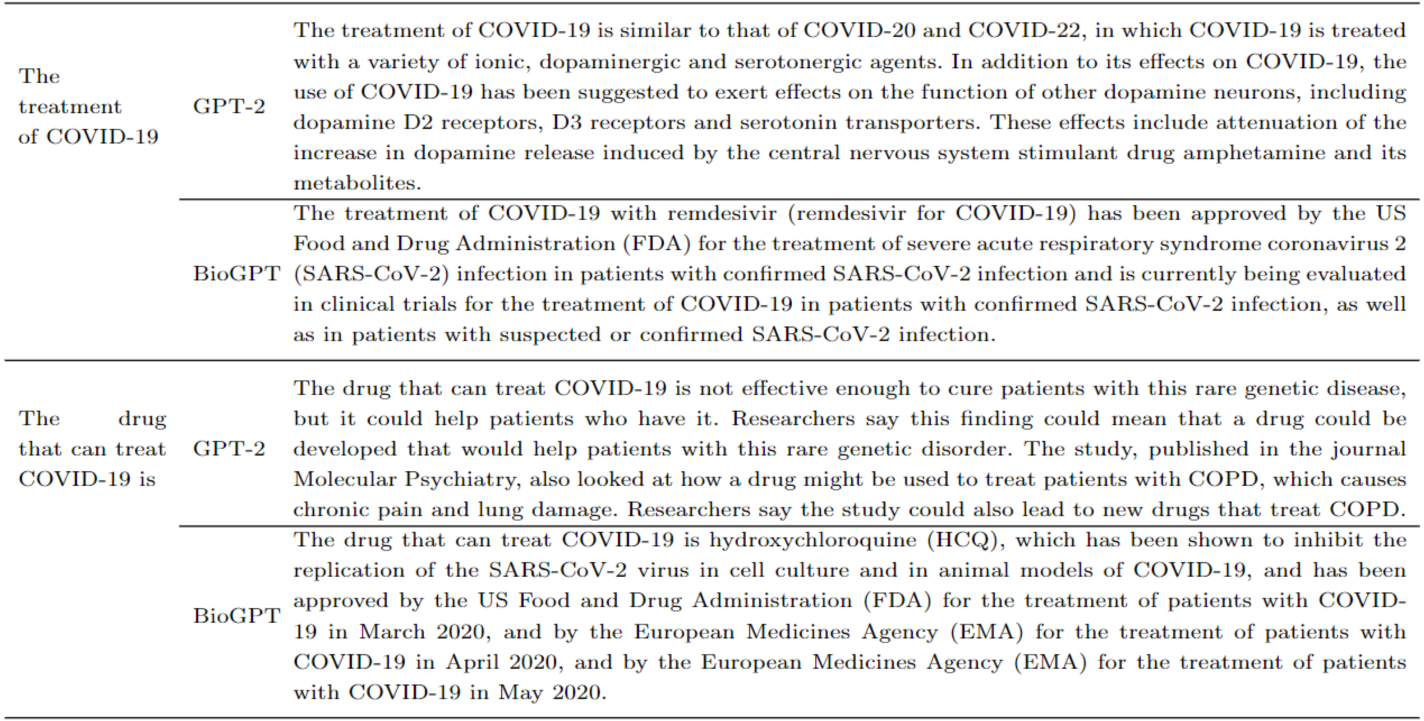
- Keywords of interest



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.

Text Generation

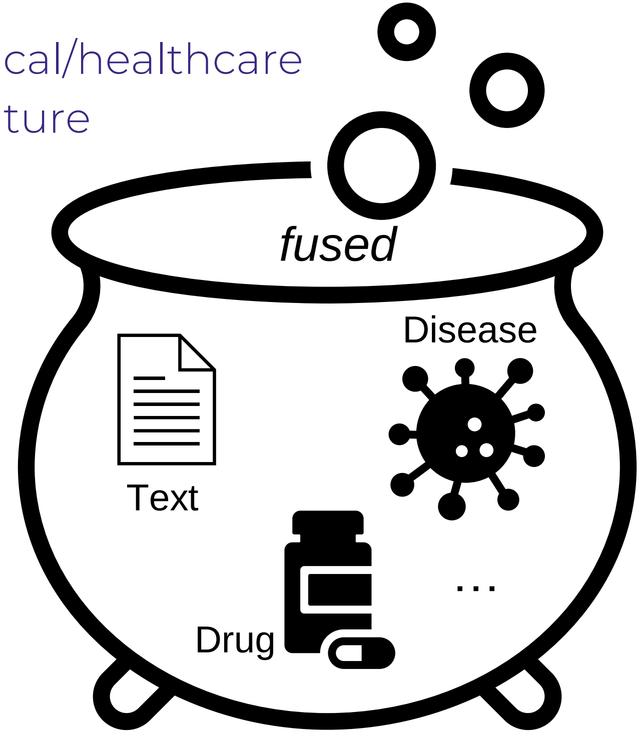
- Keywords of interest

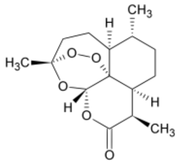


The treatment of COVID-19	GPT-2	The treatment of COVID-19 is similar to that of COVID-20 and COVID-22, in which COVID-19 is treated with a variety of ionic, dopaminergic and serotonergic agents. In addition to its effects on COVID-19, the use of COVID-19 has been suggested to exert effects on the function of other dopamine neurons, including dopamine D2 receptors, D3 receptors and serotonin transporters. These effects include attenuation of the increase in dopamine release induced by the central nervous system stimulant drug amphetamine and its metabolites.
	BioGPT	The treatment of COVID-19 with remdesivir (remdesivir for COVID-19) has been approved by the US Food and Drug Administration (FDA) for the treatment of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection in patients with confirmed SARS-CoV-2 infection and is currently being evaluated in clinical trials for the treatment of COVID-19 in patients with confirmed SARS-CoV-2 infection, as well as in patients with suspected or confirmed SARS-CoV-2 infection.
The drug that can treat COVID-19 is	GPT-2	The drug that can treat COVID-19 is not effective enough to cure patients with this rare genetic disease, but it could help patients who have it. Researchers say this finding could mean that a drug could be developed that would help patients with this rare genetic disorder. The study, published in the journal Molecular Psychiatry, also looked at how a drug might be used to treat patients with COPD, which causes chronic pain and lung damage. Researchers say the study could also lead to new drugs that treat COPD.
	BioGPT	The drug that can treat COVID-19 is hydroxychloroquine (HCQ), which has been shown to inhibit the replication of the SARS-CoV-2 virus in cell culture and in animal models of COVID-19, and has been approved by the US Food and Drug Administration (FDA) for the treatment of patients with COVID-19 in March 2020, and by the European Medicines Agency (EMA) for the treatment of patients with COVID-19 in April 2020, and by the European Medicines Agency (EMA) for the treatment of patients with COVID-19 in May 2020.

Extension: BioGPT-2

Medical/healthcare literature





Artemisinin and its derivatives represent the most important and influential class of drugs in the fight against malaria. Since …… (words from the article "Artemisinin, the Magic Drug Discovered from Traditional Chinese Medicine")

Why not put them together?

BioGPT-2 (on text and drug)

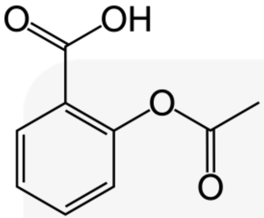
Model

BioGPT-2

48-layer Transformer
#param=1.5B

- 30M PubMed abstracts
- 30M SMILES from PubChem
- 8M pseudo parallel text-SMILES data

Aspirin

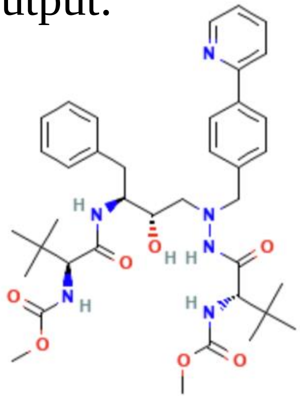


SMILES: CC(=O)Oc1ccccc1C(=O)O

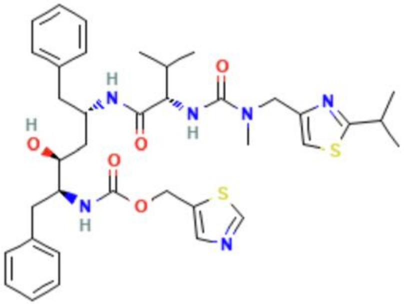
Example 1: Disease ⇒ drug

Input: generate a drug for HIV:

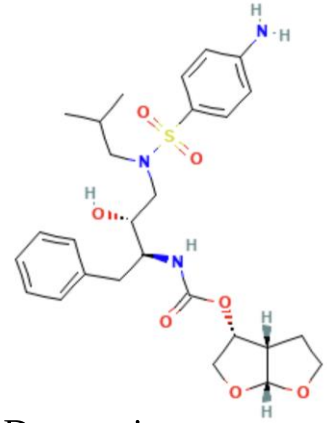
Output:



Atazanavir
Pubchem id=148192



Ritonavir
Pubchem id=392622



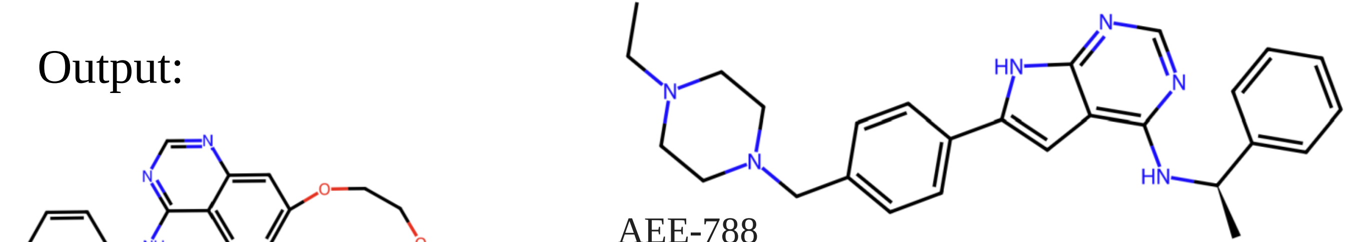
Darunavir
Pubchem id=213039

Images from
https://pubchem.ncbi.nlm.nih.gov/compound/${pubchemID}

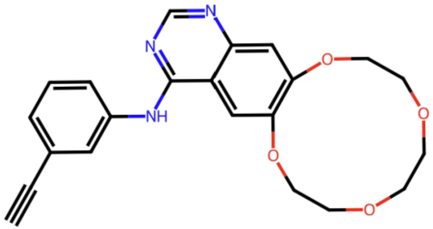
Example 2: Disease + target ⇒ drug

Input: generate drug for lung cancer targeted at EGFR.

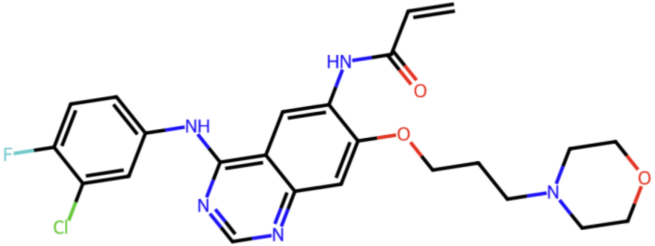
Output:



AEE-788
Pubchem id=10297043



Icotinib
https://go.drugbank.com/drugs/DB11737



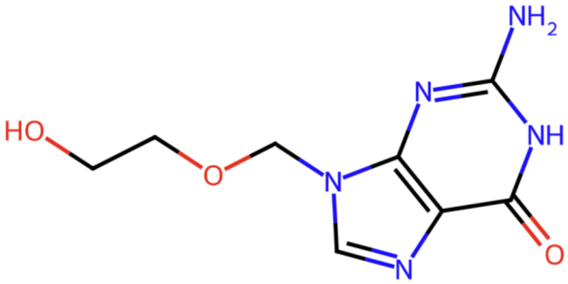
Canertinib
https://go.drugbank.com/drugs/DB05424

Example 3: Virus description ⇒ drugs

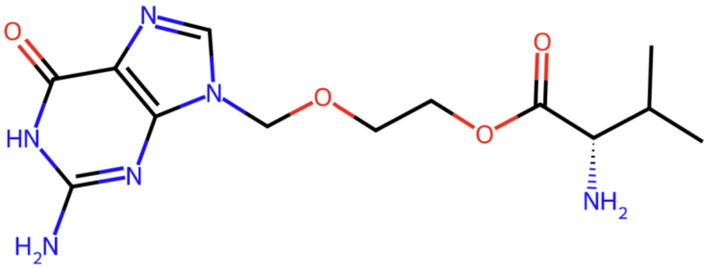
Input: The virus is closely related to the herpes simplex viruses (HSV), sharing much genome homology. The known envelope glycoproteins (gB, gC, gE, gH, gI, gK, gL) correspond with those in HSV; however, there is no equivalent of the HSV gD protein. The virus also fails to produce the LAT (latency-associated transcripts) that play an important role in establishing HSV latency (herpes simplex virus). Its virions are spherical and 180 – 200 nm in diameter. Their lipid envelope encloses the 100 nm nucleocapsid of 162 hexameric and pentameric capsomeres arranged in an icosahedral form. Its DNA is a single, linear, double-stranded molecule, 125,000 nt long. The capsid is surrounded by loosely associated proteins known collectively as the tegument; many of these proteins play critical roles in initiating the process of virus reproduction in the infected cell. The tegument is in turn covered by a lipid envelope studded with glycoproteins that are displayed on the exterior of the virion, each approximately 8 nm long. generate a drug for this virus.

// This is the description of varicella zoster virus

Output:



Acyclovir
Pubchem id=135398513



Valacyclovir
Pubchem id=135398742

BioGPT

Summary

- BioGPT: a strong generative model for biomedical domain
- Information extraction, text generation and QA

- BioGPT2: enables text to drug generation

- https://github.com/microsoft/BioGPT

- https://huggingface.co/microsoft/biogpt
- https://huggingface.co/microsoft/biogpt-large

NLP SUMMIT HEALTHCARE

Contact:
Renqian Luo, Yingce Xia, Tao Qin
{renqianluo, yingce.xia, taoqin}@microsoft.com

www.nlpsummit.org

Presented by
John Snow LABS