EECS 6895 Adv. Big Data and AI

Lecture 2: LLM and AI for Bio

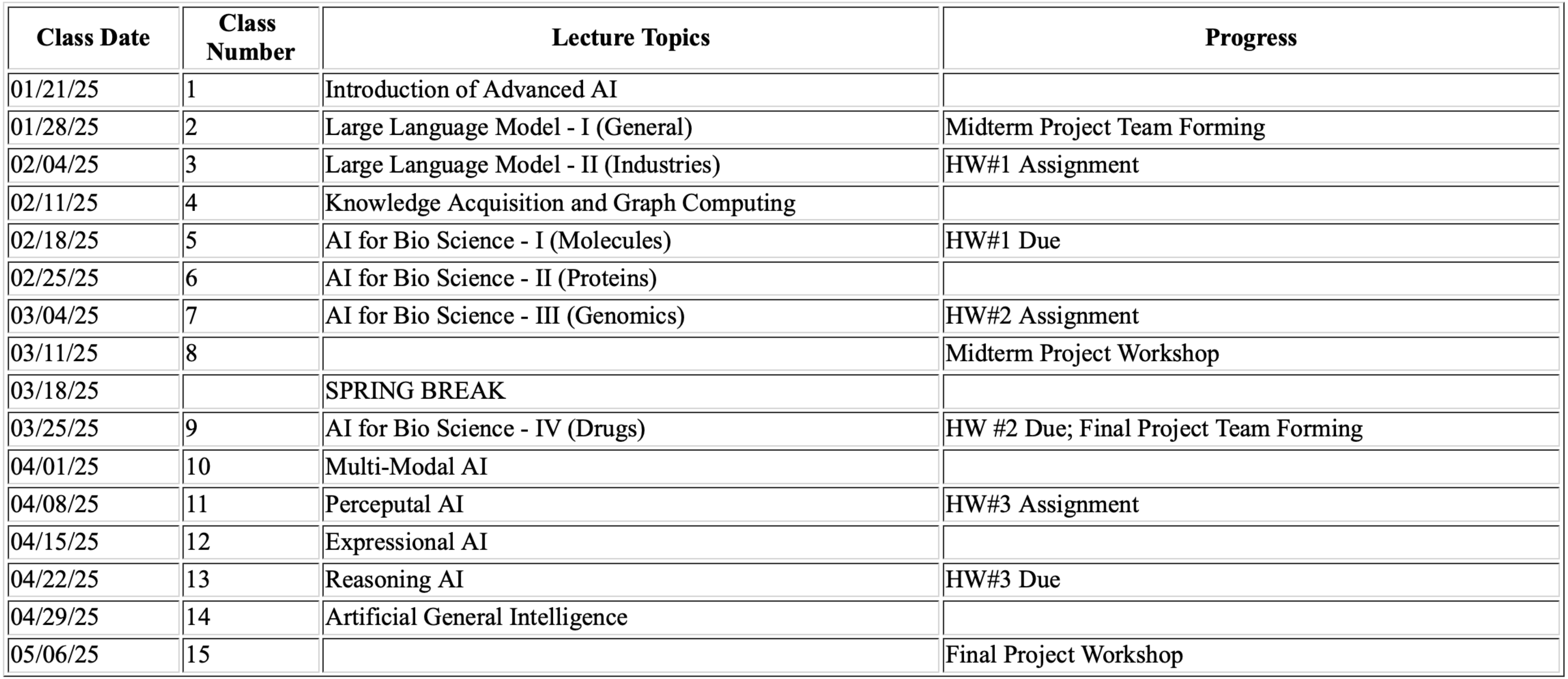
Prof. Ching-Yung Lin

Columbia University

January 28th, 2025

Recap

Course Lectures

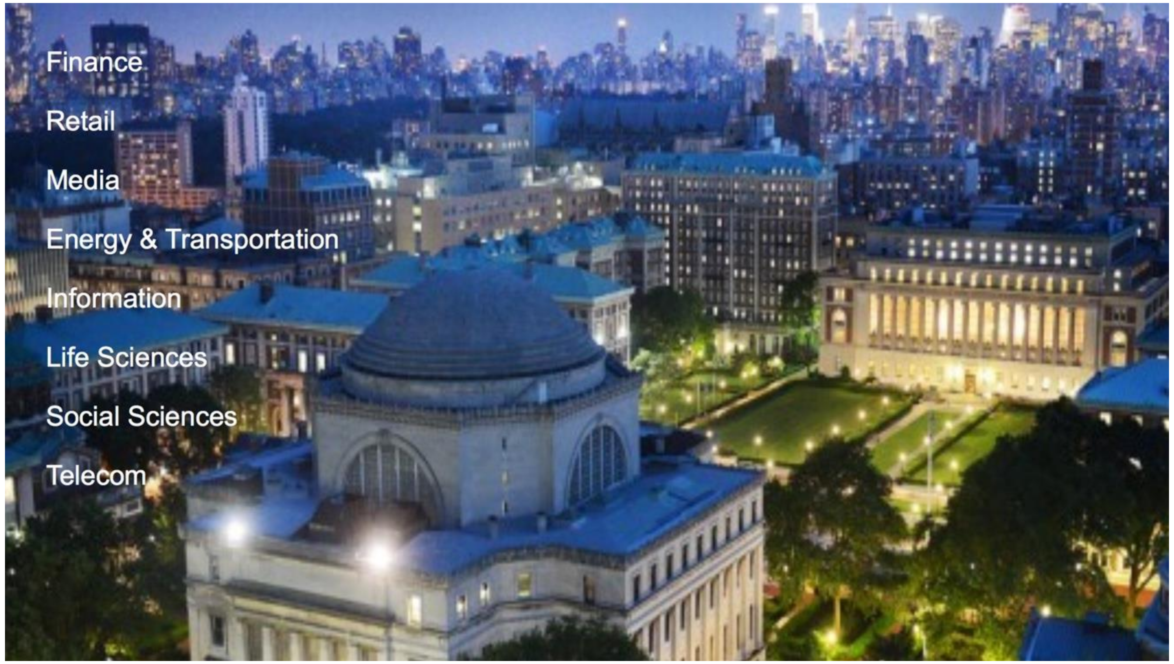


Class Date	Class Number	Lecture Topics	Progress
01/21/25	1	Introduction of Advanced AI	
01/28/25	2	Large Language Model - I (General)	Midterm Project Team Forming
02/04/25	3	Large Language Model - II (Industries)	HW#1 Assignment
02/11/25	4	Knowledge Acquisition and Graph Computing	
02/18/25	5	AI for Bio Science - I (Molecules)	HW#1 Due
02/25/25	6	AI for Bio Science - II (Proteins)	
03/04/25	7	AI for Bio Science - III (Genomics)	HW#2 Assignment
03/11/25	8		Midterm Project Workshop
03/18/25		SPRING BREAK	
03/25/25	9	AI for Bio Science - IV (Drugs)	HW #2 Due; Final Project Team Forming
04/01/25	10	Multi-Modal AI	
04/08/25	11	Perceputal AI	HW#3 Assignment
04/15/25	12	Expressional AI	
04/22/25	13	Reasoning AI	HW#3 Due
04/29/25	14	Artificial General Intelligence	
05/06/25	15		Final Project Workshop

*Lectures: 19:00 – 20:20 *Group Presentations: 20:30 – 21:30

Course Information and TAs

- Professor Lin:
  - Office Hours and Location: By appointment (500 Fifth Ave., Suite 2420, New York, NY 10110)
  - Contact: c.lin@columbia.edu
- TAs:
  - Zelin Yu (zy2489)
  - Likhith Ayinala (la3073)



- Website (Materials):

○ http://www.ee.columbia.edu/~cylin/course/bigdata/

Course Grading

- Midterm Project (40%):
  - Advanced A.I. Assistants for Financial Industry and Healthcare Industry.

- Final Project (40%):
  - Advanced A.I. Technologies; *or*
  - Advanced A.I. for Bio Science.

- 3 Homeworks (15%):
  - AI Assistant, AI for Bio Science, and Advanced AI Technology

- Course Participation (5%) :
  - Attendance and Discussions

Midterm Project

- Each team is composed of 3 people.

- Choose among these areas. Each area has two teams.

  - Financial Industry:
    - Tax Advisor
    - Insurance Advisor
    - Private Banking Specialist
    - Commercial Banking Specialist
    - Fund Manager / M&A Specialist

  - Healthcare Industry:
    - Nurse
    - Doctor
    - Nutritionist
    - Pharmacist
    - Radiologist

Midterm Project – To-Do by 1/28/25

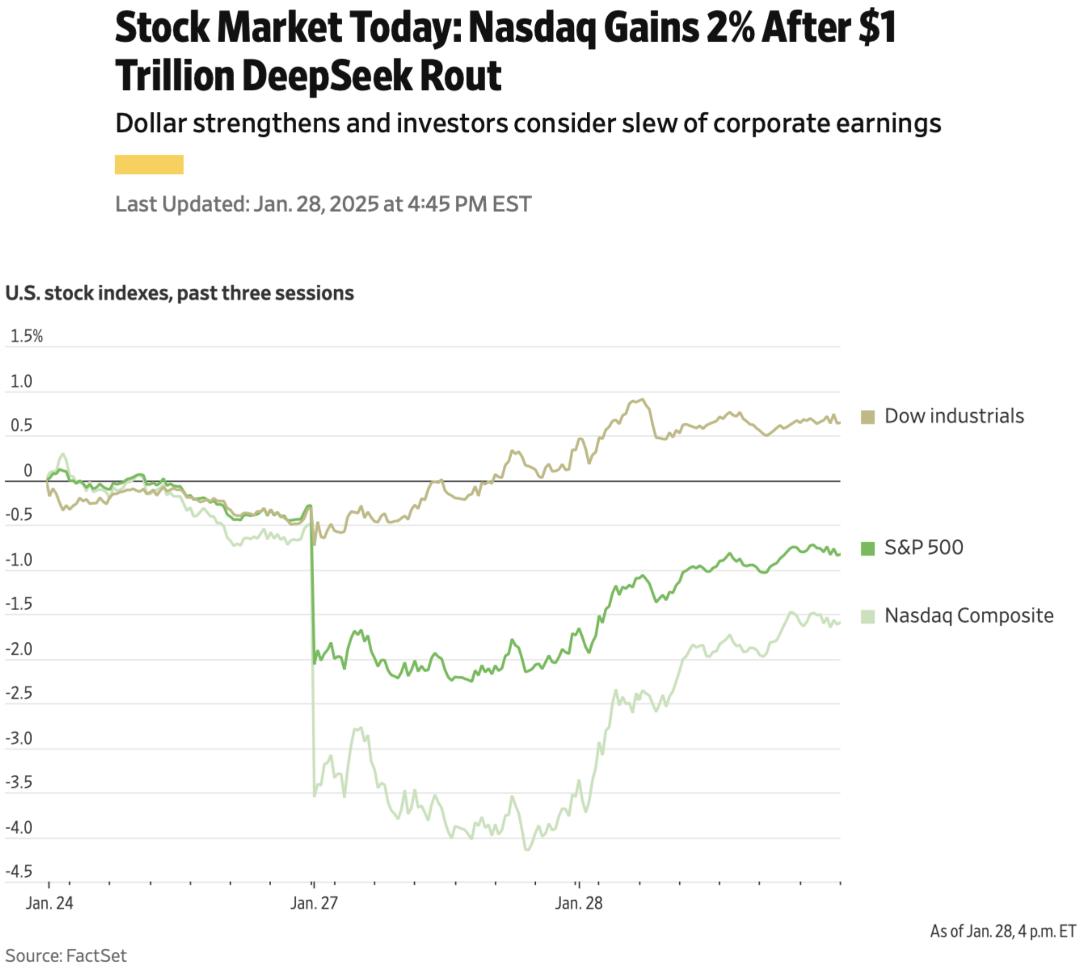
- Sign up the team sheet and project choice – to be shared by TA.

- Each team prepare a 3-min presentation to discuss -- voluntarily.

  - The Certificate Challenge
  - Initial Thought of Knowledges and Capabilities of the AI Assistant to be included

What’s happening?

Stock market plunge and rebound because of AI



Stock Market Today: Nasdaq Gains 2% After $1 Trillion DeepSeek Rout

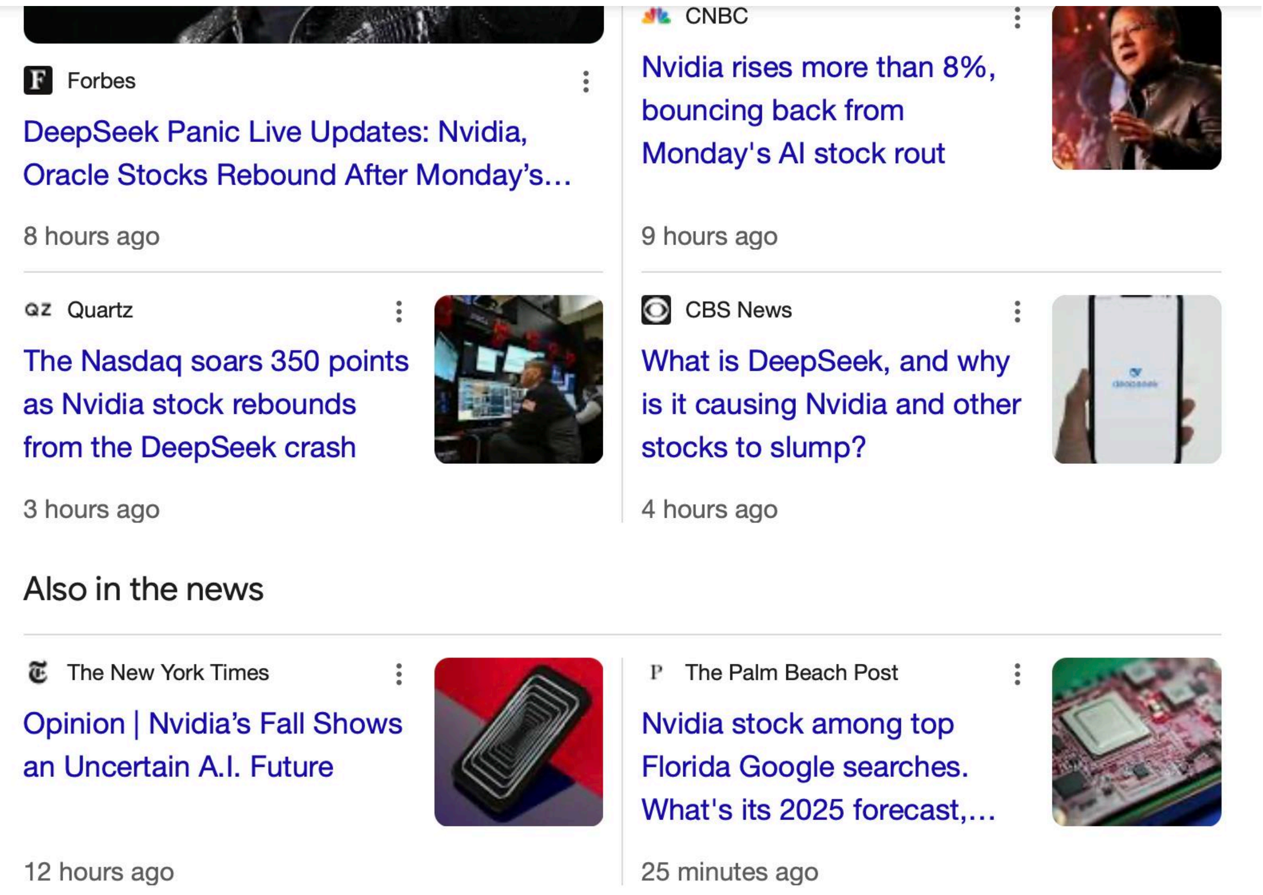
Dollar strengthens and investors consider slew of corporate earnings

Last Updated: Jan. 28, 2025 at 4:45 PM EST

U.S. stock indexes, past three sessions

As of Jan. 28, 4 p.m. ET

Source: FactSet



Forbes

DeepSeek Panic Live Updates: Nvidia, Oracle Stocks Rebound After Monday's…

8 hours ago

CNBC

Nvidia rises more than 8%, bouncing back from Monday's AI stock rout

9 hours ago

Quartz

The Nasdaq soars 350 points as Nvidia stock rebounds from the DeepSeek crash

3 hours ago

CBS News

What is DeepSeek, and why is it causing Nvidia and other stocks to slump?

4 hours ago

Also in the news

The New York Times

Opinion | Nvidia's Fall Shows an Uncertain A.I. Future

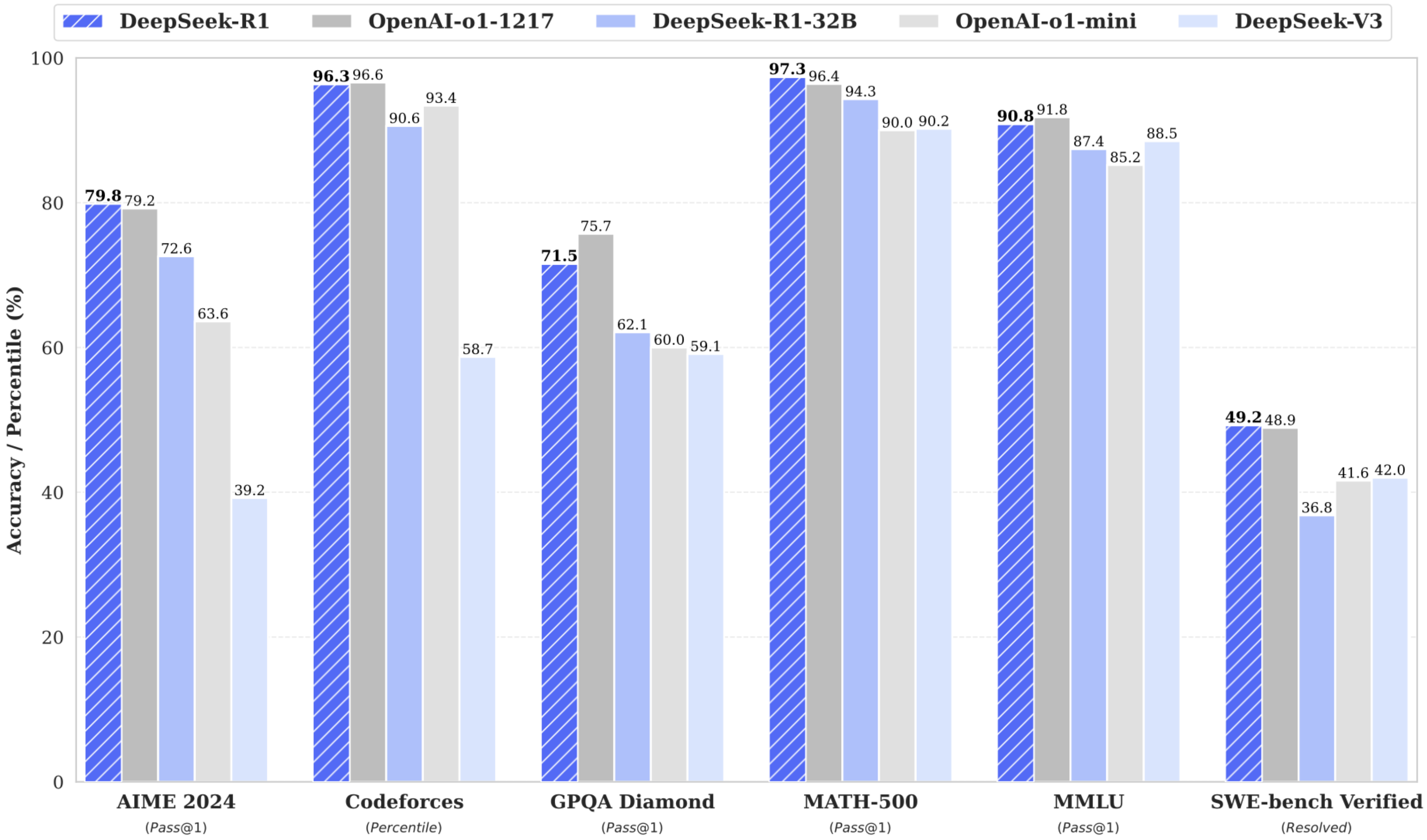
12 hours ago

The Palm Beach Post

Nvidia stock among top Florida Google searches. What's its 2025 forecast,…

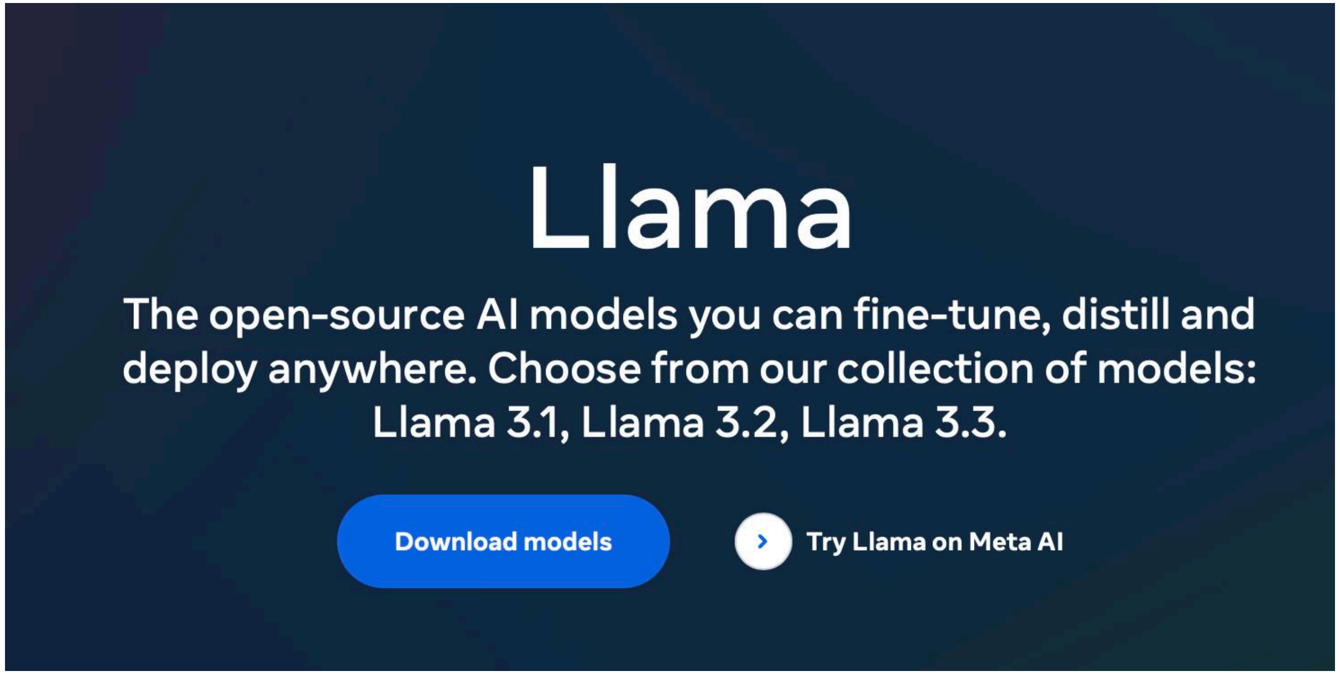
25 minutes ago

Developing towards AGI



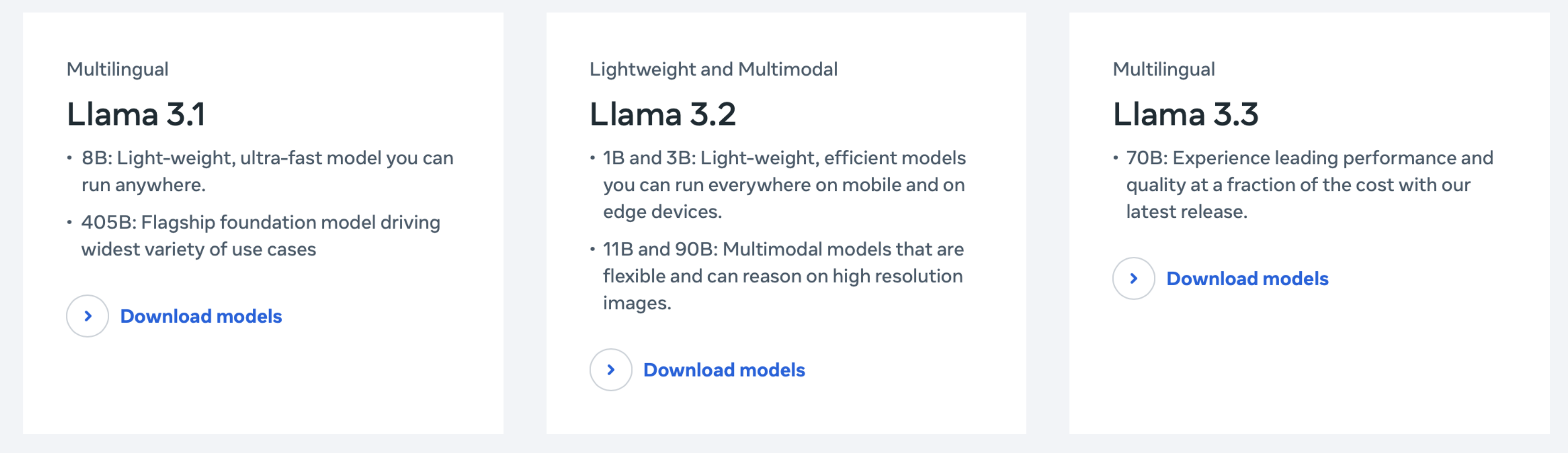
DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning https://arxiv.org/pdf/2501.12948

Llama



Llama

The open-source AI models you can fine-tune, distill and deploy anywhere. Choose from our collection of models: Llama 3.1, Llama 3.2, Llama 3.3.



Multilingual

Llama 3.1

- 8B: Light-weight, ultra-fast model you can run anywhere.
- 405B: Flagship foundation model driving widest variety of use cases

Lightweight and Multimodal

Llama 3.2

- 1B and 3B: Light-weight, efficient models you can run everywhere on mobile and on edge devices.
- 11B and 90B: Multimodal models that are flexible and can reason on high resolution images.

Multilingual

Llama 3.3

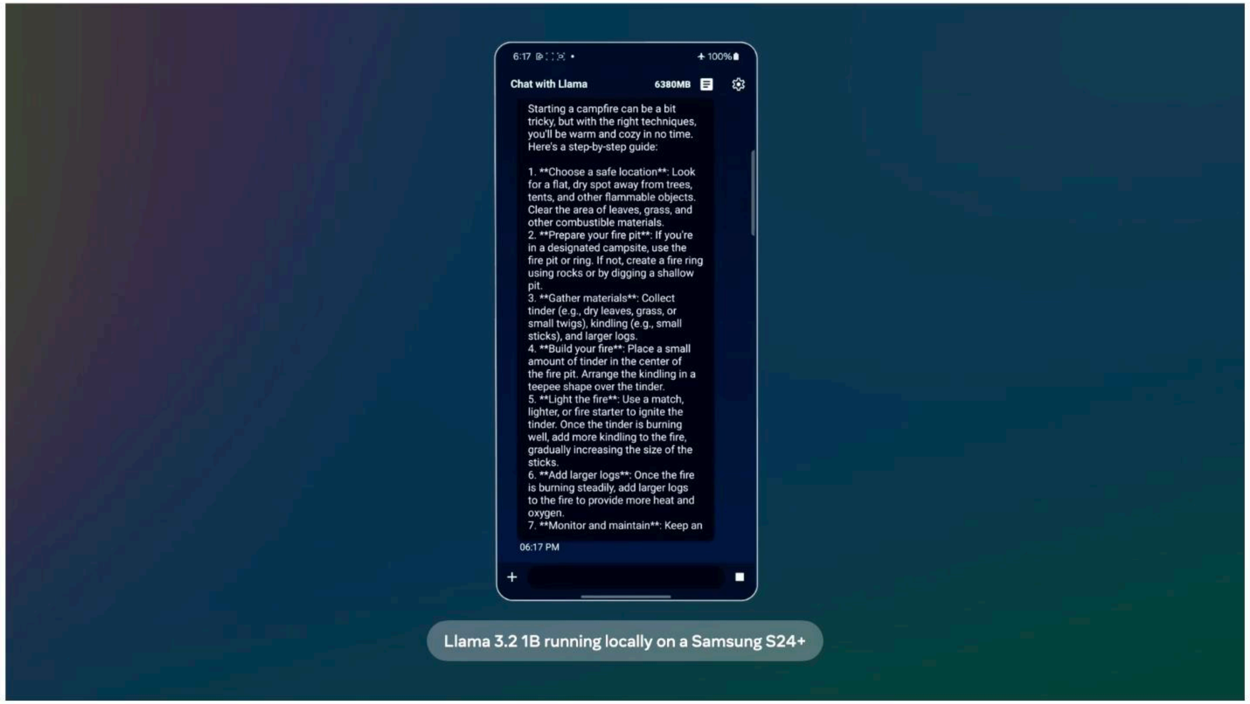
- 70B: Experience leading performance and quality at a fraction of the cost with our latest release.

Llama 3.2

Explore the new capabilities of Llama 3.2

The Llama 3.2 lightweight models enable Llama to run on phones, tablets, and edge devices. View the video to see Llama running on phone. To see how this demo was implemented, check out **the example code** from ExecuTorch.

› **Learn more**

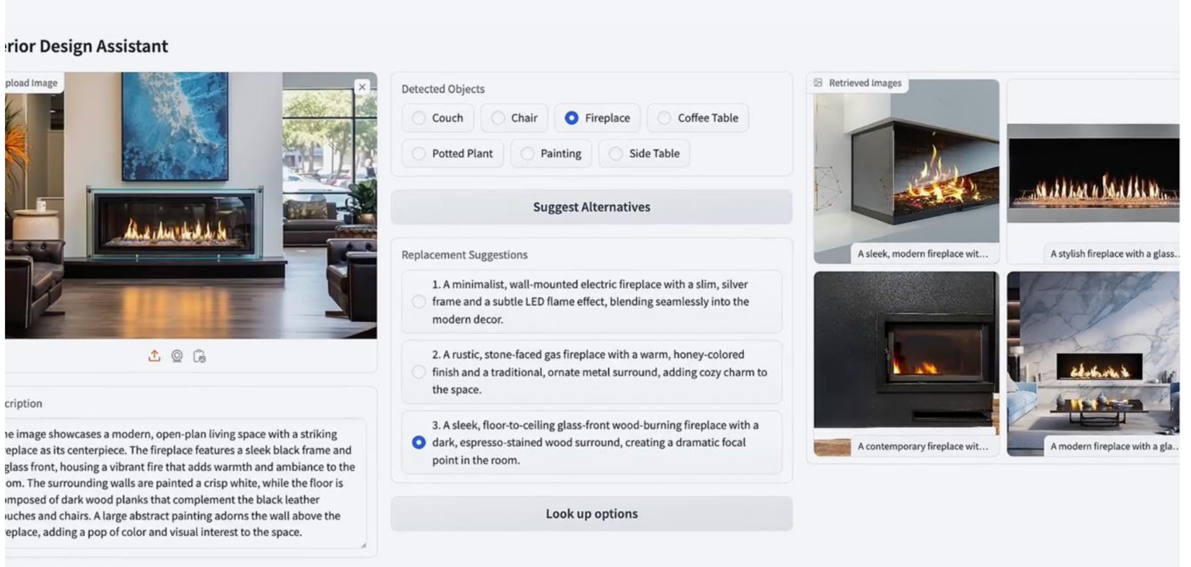


Starting a campfire can be a bit tricky, but with the right techniques, you'll be warm and cozy in no time. Here's a step-by-step guide:

1. **Choose a safe location**: Look for a flat, dry spot away from trees, tents, and other flammable objects. Clear the area of leaves, grass, and other combustible materials.
2. **Prepare your fire pit**: If you're in a designated campsite, use the fire pit or ring. If not, create a fire ring using rocks or by digging a shallow pit.
3. **Gather materials**: Collect tinder (e.g., dry leaves, grass, or small twigs), kindling (e.g., small sticks), and larger logs.
4. **Build your fire**: Place a small amount of tinder in the center of the fire pit. Arrange the kindling in a teepee shape over the tinder.
5. **Light the fire**: Use a match, lighter, or fire starter to ignite the tinder. Once the tinder is burning well, add more kindling to the fire, gradually increasing the size of the sticks.
6. **Add larger logs**: Once the fire is burning steadily, add larger logs to the fire to provide more heat and oxygen.
7. **Monitor and maintain**: Keep an

06:17 PM

Llama 3.2 1B running locally on a Samsung S24+



Detected Objects

○ Couch ○ Chair ● Fireplace ○ Coffee Table ○ Potted Plant ○ Painting ○ Side Table

Replacement Suggestions

○ 1. A minimalist, wall-mounted electric fireplace with a slim, silver frame and a subtle LED flame effect, blending seamlessly into the modern decor.

○ 2. A rustic, stone-faced gas fireplace with a warm, honey-colored finish and a traditional, ornate metal surround, adding cozy charm to the space.

● 3. A sleek, floor-to-ceiling glass-front wood-burning fireplace with a dark, espresso-stained wood surround, creating a dramatic focal point in the room.

cription

ne image showcases a modern, open-plan living space with a striking eplace as its centerpiece. The fireplace features a sleek black frame and glass front, housing a vibrant fire that adds warmth and ambiance to the om. The surrounding walls are painted a crisp white, while the floor is mposed of dark wood planks that complement the black leather uches and chairs. A large abstract painting adorns the wall above the eplace, adding a pop of color and visual interest to the space.

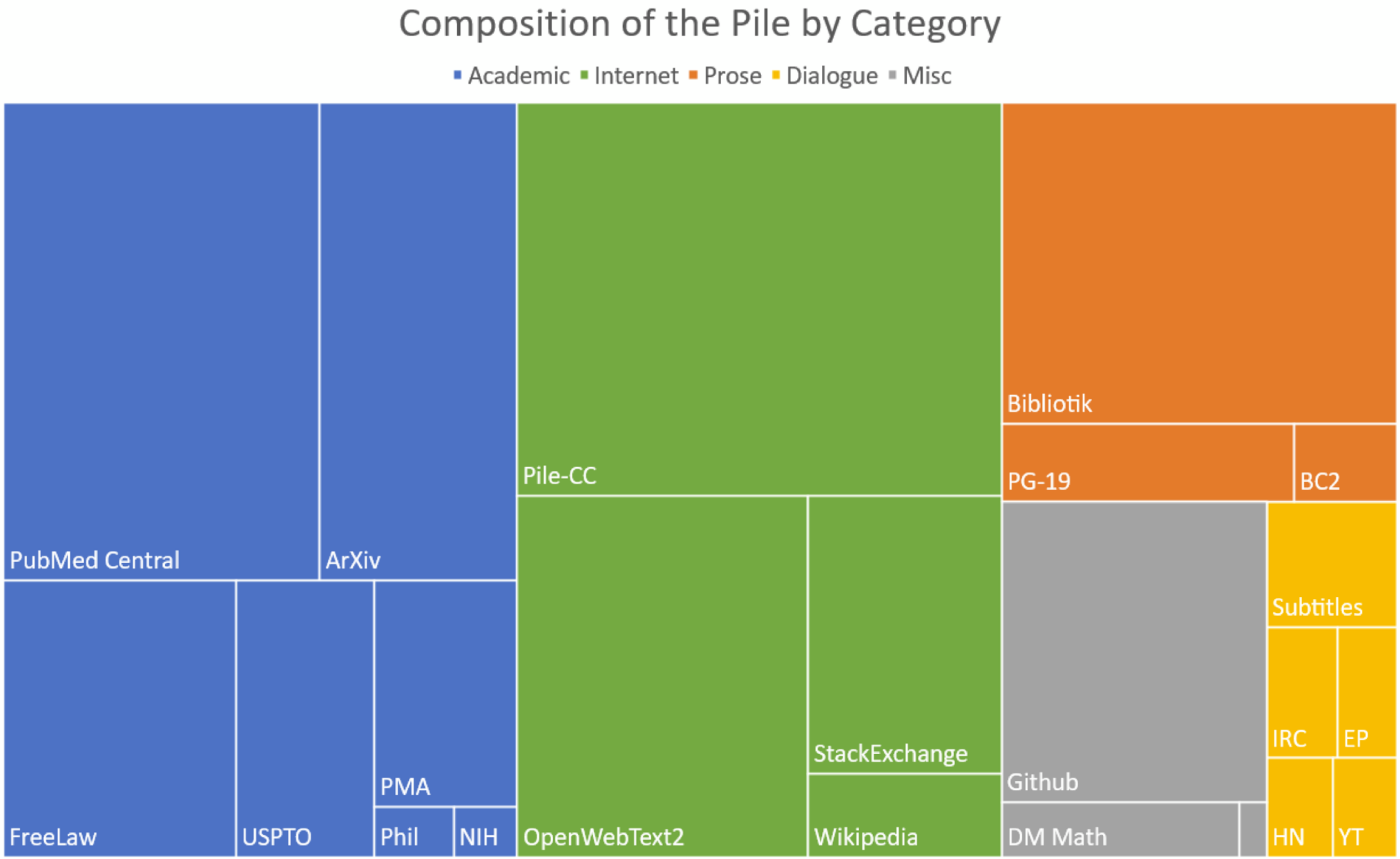
A sleek, modern fireplace wit...

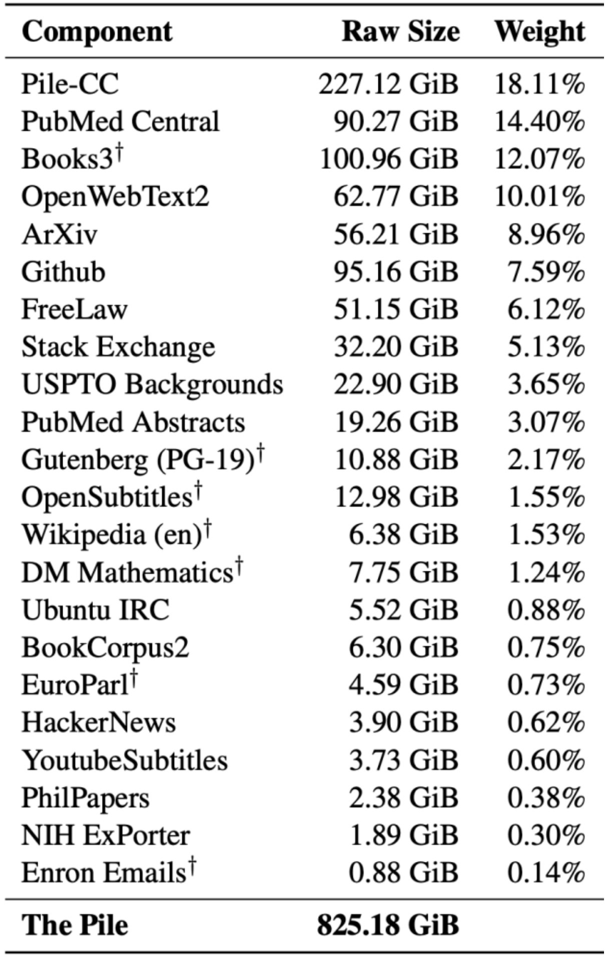
A stylish fireplace with a glass...

A contemporary fireplace wit...

A modern fireplace with a gla...

Pretrained models using the Pile dataset





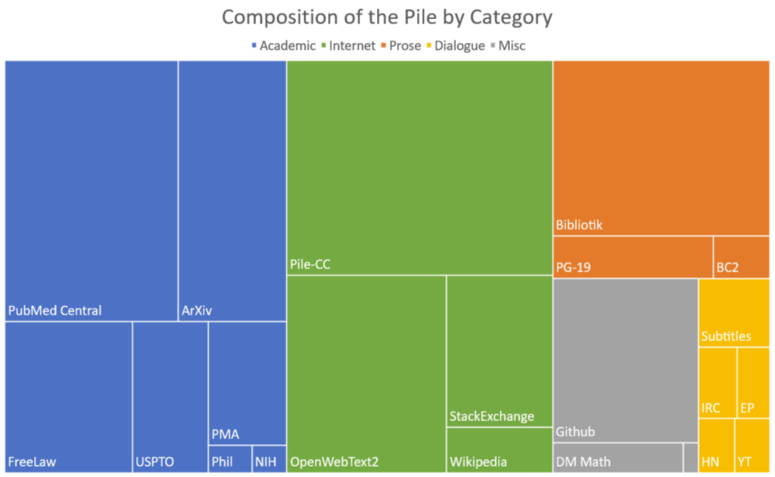
Component	Raw Size	Weight
Pile-CC	227.12 GiB	18.11%
PubMed Central	90.27 GiB	14.40%
Books3†	100.96 GiB	12.07%
OpenWebText2	62.77 GiB	10.01%
ArXiv	56.21 GiB	8.96%
Github	95.16 GiB	7.59%
FreeLaw	51.15 GiB	6.12%
Stack Exchange	32.20 GiB	5.13%
USPTO Backgrounds	22.90 GiB	3.65%
PubMed Abstracts	19.26 GiB	3.07%
Gutenberg (PG-19)†	10.88 GiB	2.17%
OpenSubtitles†	12.98 GiB	1.55%
Wikipedia (en)†	6.38 GiB	1.53%
DM Mathematics†	7.75 GiB	1.24%
Ubuntu IRC	5.52 GiB	0.88%
BookCorpus2	6.30 GiB	0.75%
EuroParl†	4.59 GiB	0.73%
HackerNews	3.90 GiB	0.62%
YoutubeSubtitles	3.73 GiB	0.60%
PhilPapers	2.38 GiB	0.38%
NIH ExPorter	1.89 GiB	0.30%
Enron Emails†	0.88 GiB	0.14%
The Pile	**825.18 GiB**	

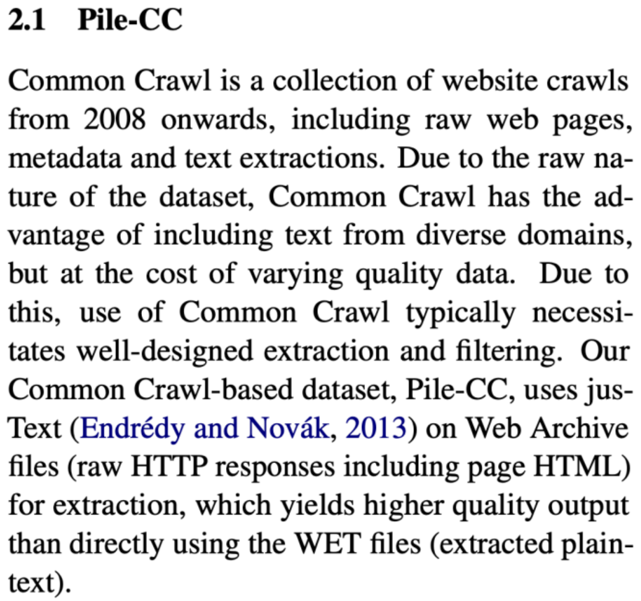
https://pile.eleuther.ai/paper.pdf

The Pile dataset - I

Columbia University

Composition of the Pile by Category

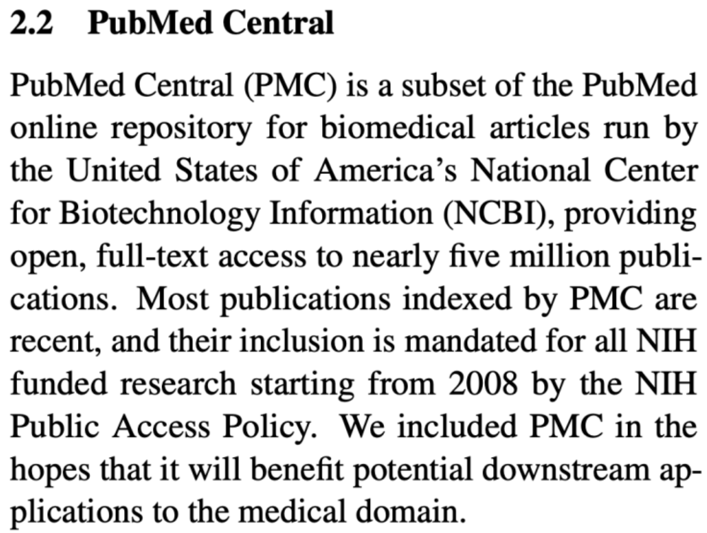




2.1 Pile-CC

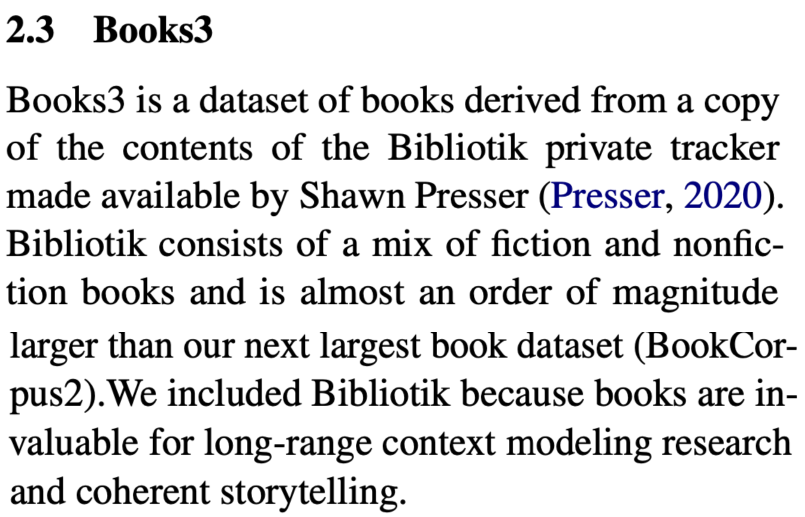
Common Crawl is a collection of website crawls from 2008 onwards, including raw web pages, metadata and text extractions. Due to the raw nature of the dataset, Common Crawl has the advantage of including text from diverse domains, but at the cost of varying quality data. Due to this, use of Common Crawl typically necessitates well-designed extraction and filtering. Our Common Crawl-based dataset, Pile-CC, uses jusText (Endrédy and Novák, 2013) on Web Archive files (raw HTTP responses including page HTML) for extraction, which yields higher quality output than directly using the WET files (extracted plaintext).

https://pile.eleuther.ai/paper.pdf



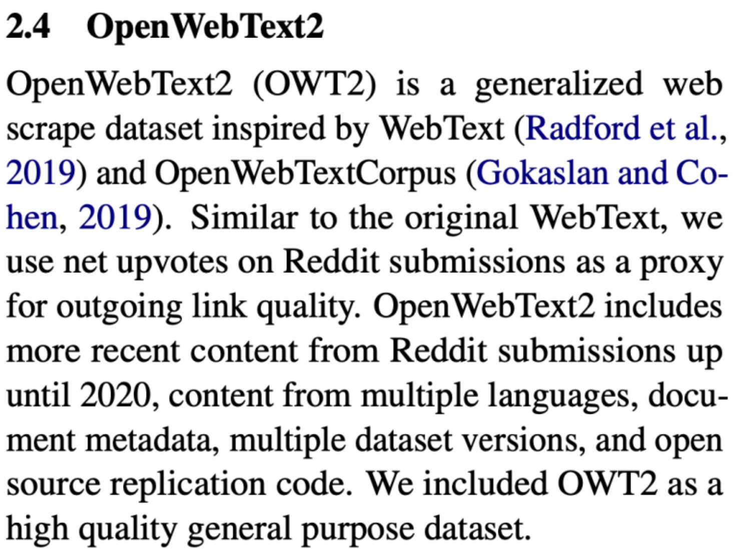
2.2 PubMed Central

PubMed Central (PMC) is a subset of the PubMed online repository for biomedical articles run by the United States of America's National Center for Biotechnology Information (NCBI), providing open, full-text access to nearly five million publications. Most publications indexed by PMC are recent, and their inclusion is mandated for all NIH funded research starting from 2008 by the NIH Public Access Policy. We included PMC in the hopes that it will benefit potential downstream applications to the medical domain.



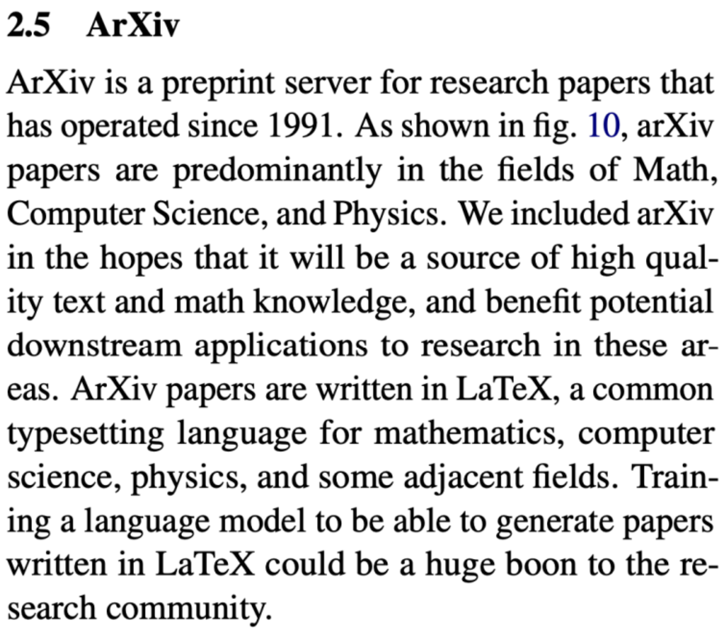
2.3 Books3

Books3 is a dataset of books derived from a copy of the contents of the Bibliotik private tracker made available by Shawn Presser (Presser, 2020). Bibliotik consists of a mix of fiction and nonfiction books and is almost an order of magnitude larger than our next largest book dataset (BookCorpus2).We included Bibliotik because books are invaluable for long-range context modeling research and coherent storytelling.



2.4 OpenWebText2

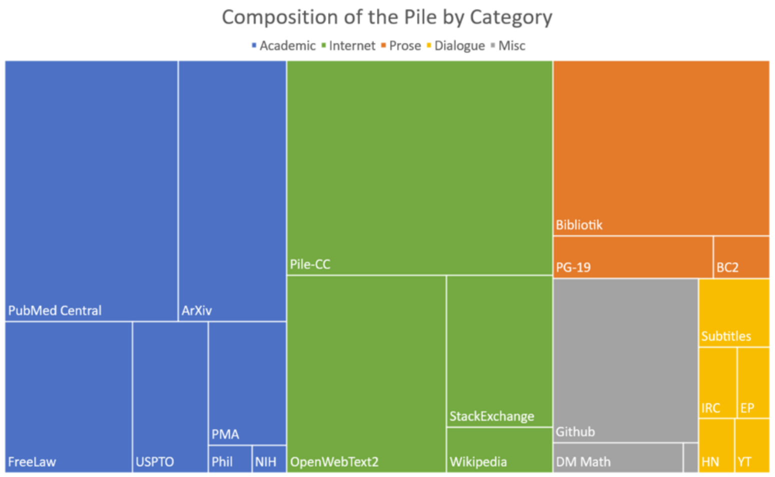
OpenWebText2 (OWT2) is a generalized web scrape dataset inspired by WebText (Radford et al., 2019) and OpenWebTextCorpus (Gokaslan and Cohen, 2019). Similar to the original WebText, we use net upvotes on Reddit submissions as a proxy for outgoing link quality. OpenWebText2 includes more recent content from Reddit submissions up until 2020, content from multiple languages, document metadata, multiple dataset versions, and open source replication code. We included OWT2 as a high quality general purpose dataset.

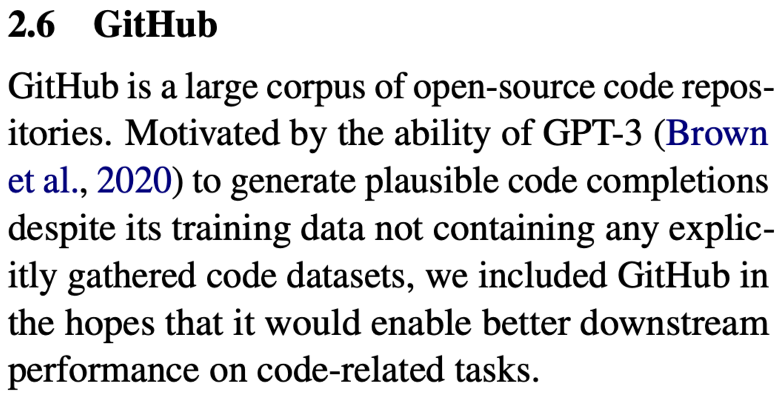


2.5 ArXiv

ArXiv is a preprint server for research papers that has operated since 1991. As shown in fig. 10, arXiv papers are predominantly in the fields of Math, Computer Science, and Physics. We included arXiv in the hopes that it will be a source of high quality text and math knowledge, and benefit potential downstream applications to research in these areas. ArXiv papers are written in LaTeX, a common typesetting language for mathematics, computer science, physics, and some adjacent fields. Training a language model to be able to generate papers written in LaTeX could be a huge boon to the research community.

The Pile dataset - II

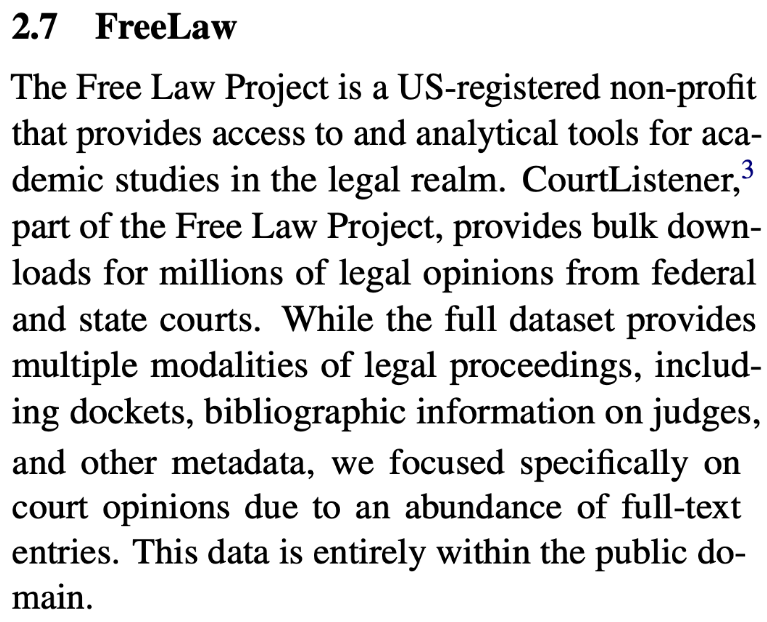




2.6 GitHub

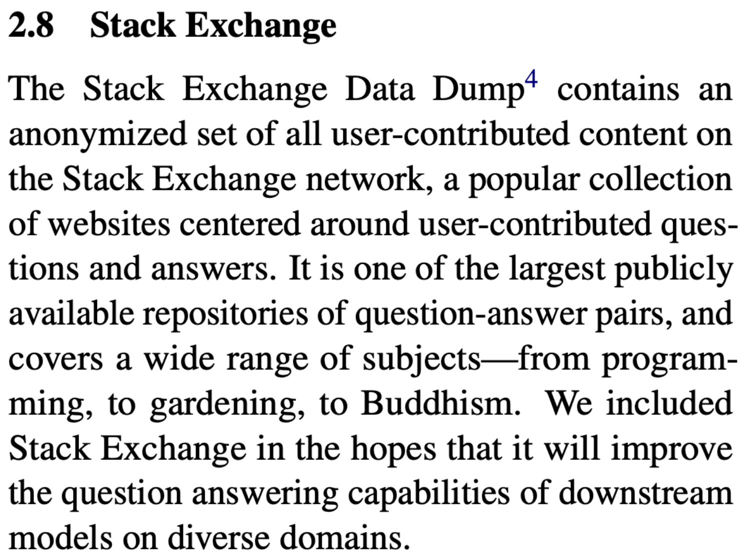
GitHub is a large corpus of open-source code repositories. Motivated by the ability of GPT-3 (Brown et al., 2020) to generate plausible code completions despite its training data not containing any explicitly gathered code datasets, we included GitHub in the hopes that it would enable better downstream performance on code-related tasks.

https://pile.eleuther.ai/paper.pdf



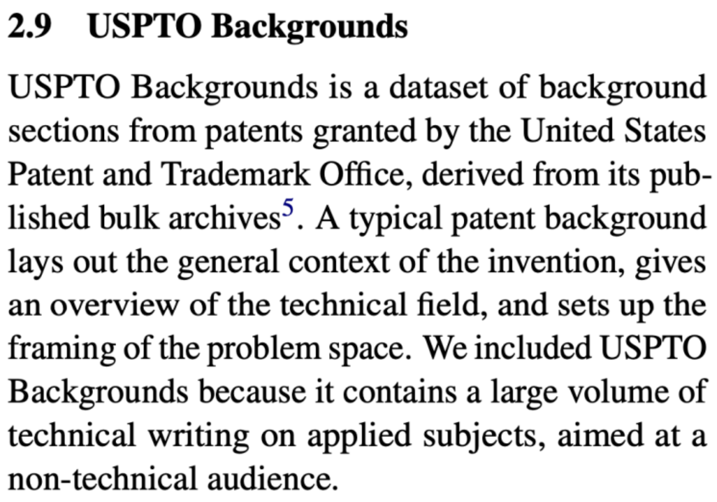
2.7 FreeLaw

The Free Law Project is a US-registered non-profit that provides access to and analytical tools for academic studies in the legal realm. CourtListener,[3] part of the Free Law Project, provides bulk downloads for millions of legal opinions from federal and state courts. While the full dataset provides multiple modalities of legal proceedings, including dockets, bibliographic information on judges, and other metadata, we focused specifically on court opinions due to an abundance of full-text entries. This data is entirely within the public domain.



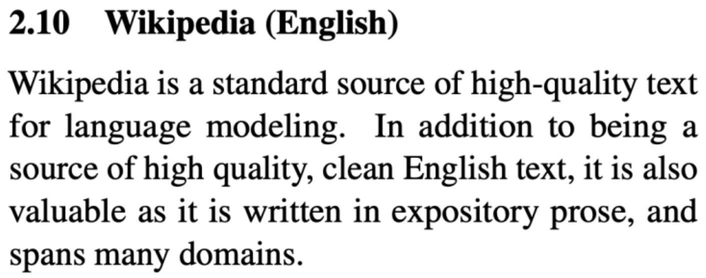
2.8 Stack Exchange

The Stack Exchange Data Dump[4] contains an anonymized set of all user-contributed content on the Stack Exchange network, a popular collection of websites centered around user-contributed questions and answers. It is one of the largest publicly available repositories of question-answer pairs, and covers a wide range of subjects—from programming, to gardening, to Buddhism. We included Stack Exchange in the hopes that it will improve the question answering capabilities of downstream models on diverse domains.



2.9 USPTO Backgrounds

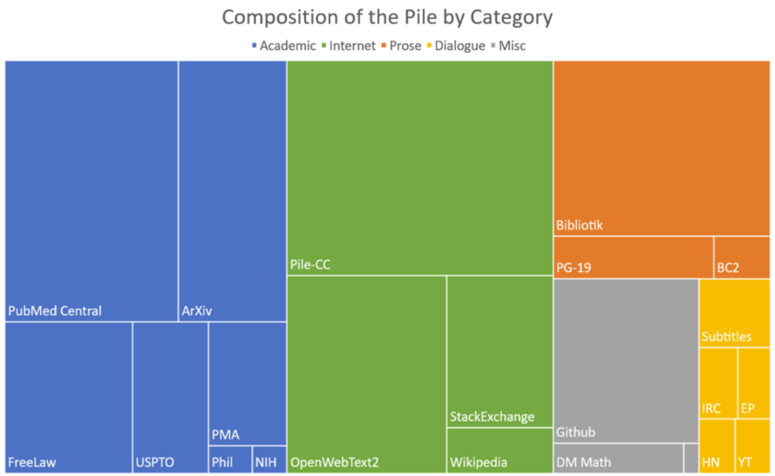
USPTO Backgrounds is a dataset of background sections from patents granted by the United States Patent and Trademark Office, derived from its published bulk archives[5]. A typical patent background lays out the general context of the invention, gives an overview of the technical field, and sets up the framing of the problem space. We included USPTO Backgrounds because it contains a large volume of technical writing on applied subjects, aimed at a non-technical audience.

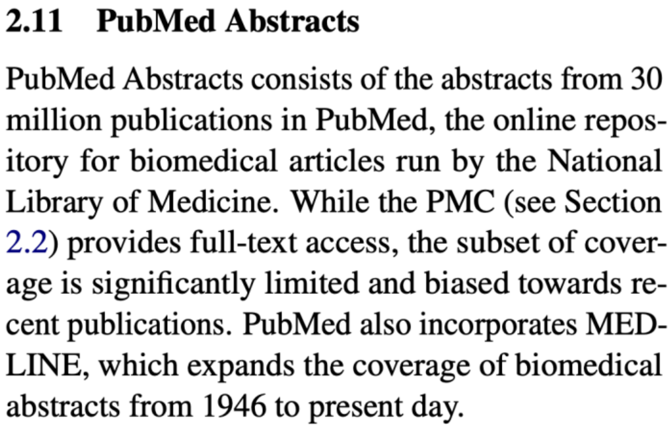


2.10 Wikipedia (English)

Wikipedia is a standard source of high-quality text for language modeling. In addition to being a source of high quality, clean English text, it is also valuable as it is written in expository prose, and spans many domains.

The Pile dataset - III

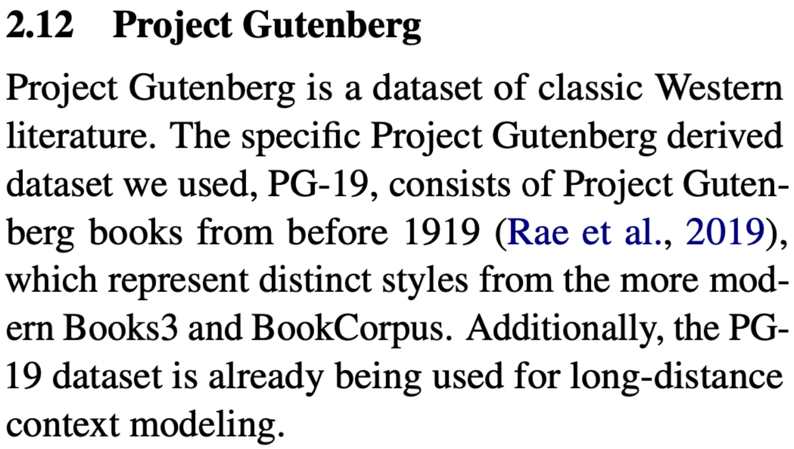




2.11 PubMed Abstracts

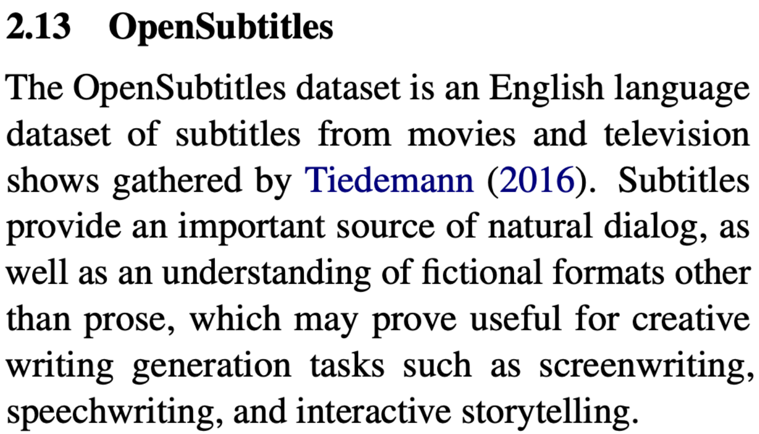
PubMed Abstracts consists of the abstracts from 30 million publications in PubMed, the online repository for biomedical articles run by the National Library of Medicine. While the PMC (see Section 2.2) provides full-text access, the subset of coverage is significantly limited and biased towards recent publications. PubMed also incorporates MEDLINE, which expands the coverage of biomedical abstracts from 1946 to present day.

https://pile.eleuther.ai/paper.pdf



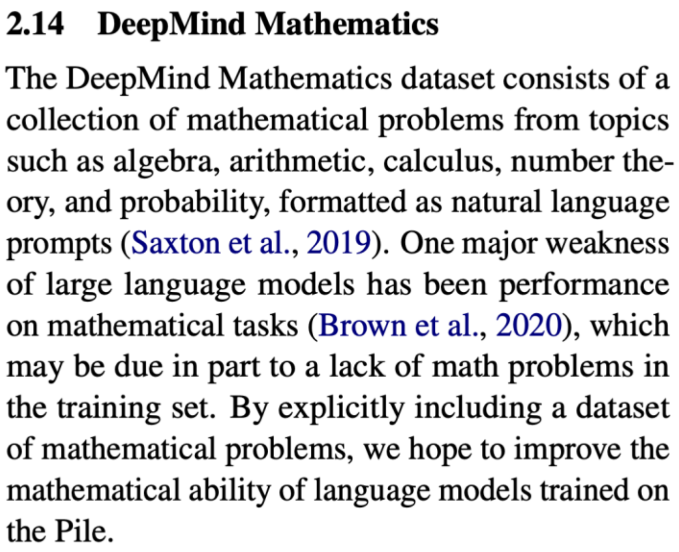
2.12 Project Gutenberg

Project Gutenberg is a dataset of classic Western literature. The specific Project Gutenberg derived dataset we used, PG-19, consists of Project Gutenberg books from before 1919 (Rae et al., 2019), which represent distinct styles from the more modern Books3 and BookCorpus. Additionally, the PG-19 dataset is already being used for long-distance context modeling.



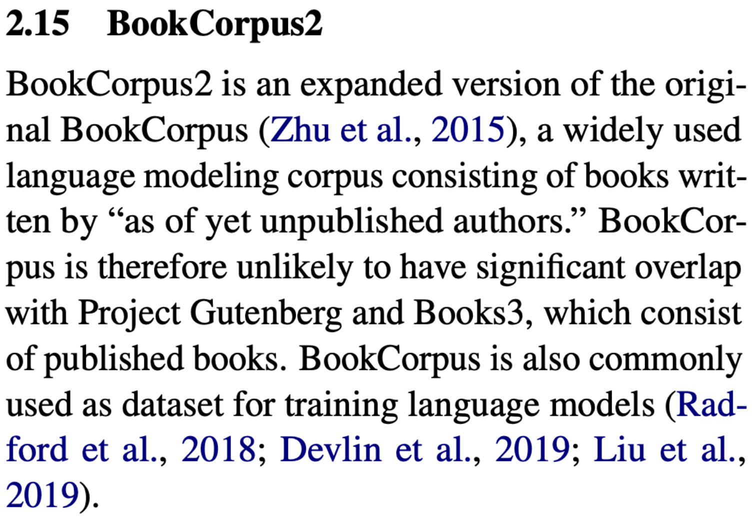
2.13 OpenSubtitles

The OpenSubtitles dataset is an English language dataset of subtitles from movies and television shows gathered by Tiedemann (2016). Subtitles provide an important source of natural dialog, as well as an understanding of fictional formats other than prose, which may prove useful for creative writing generation tasks such as screenwriting, speechwriting, and interactive storytelling.



2.14 DeepMind Mathematics

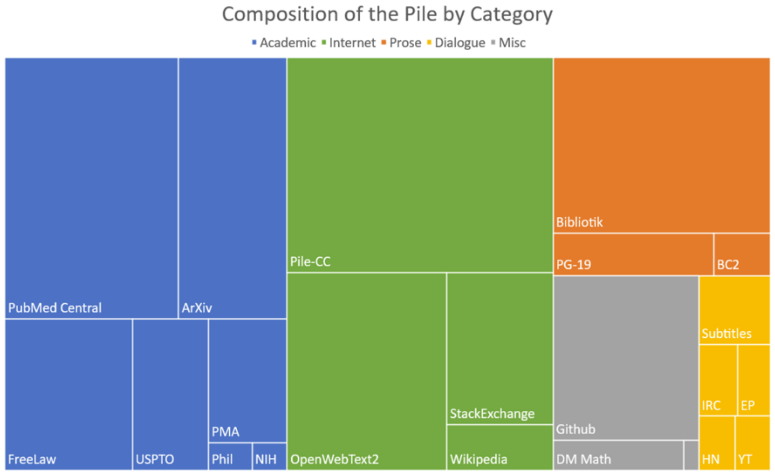
The DeepMind Mathematics dataset consists of a collection of mathematical problems from topics such as algebra, arithmetic, calculus, number theory, and probability, formatted as natural language prompts (Saxton et al., 2019). One major weakness of large language models has been performance on mathematical tasks (Brown et al., 2020), which may be due in part to a lack of math problems in the training set. By explicitly including a dataset of mathematical problems, we hope to improve the mathematical ability of language models trained on the Pile.

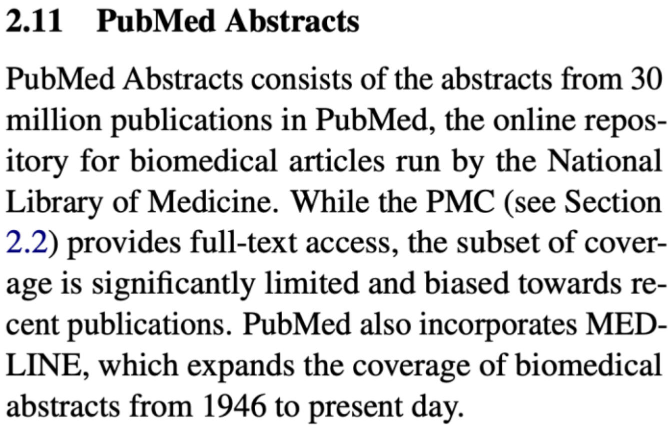


2.15 BookCorpus2

BookCorpus2 is an expanded version of the original BookCorpus (Zhu et al., 2015), a widely used language modeling corpus consisting of books written by "as of yet unpublished authors." BookCorpus is therefore unlikely to have significant overlap with Project Gutenberg and Books3, which consist of published books. BookCorpus is also commonly used as dataset for training language models (Radford et al., 2018; Devlin et al., 2019; Liu et al., 2019).

The Pile dataset - IV

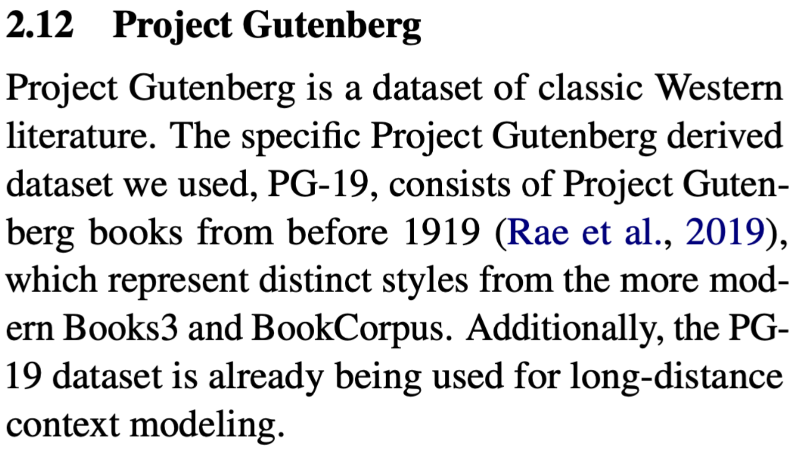




2.11 PubMed Abstracts

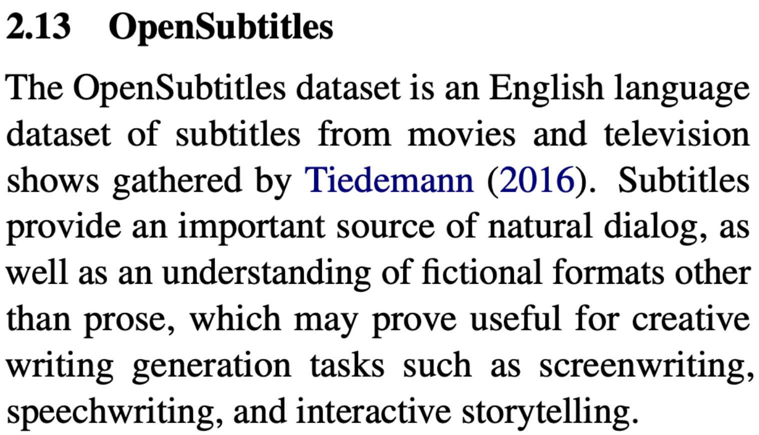
PubMed Abstracts consists of the abstracts from 30 million publications in PubMed, the online repository for biomedical articles run by the National Library of Medicine. While the PMC (see Section 2.2) provides full-text access, the subset of coverage is significantly limited and biased towards recent publications. PubMed also incorporates MEDLINE, which expands the coverage of biomedical abstracts from 1946 to present day.

https://pile.eleuther.ai/paper.pdf



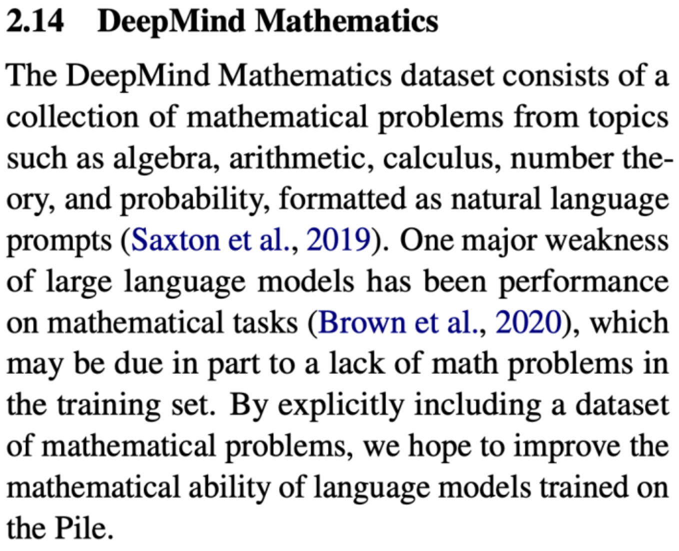
2.12 Project Gutenberg

Project Gutenberg is a dataset of classic Western literature. The specific Project Gutenberg derived dataset we used, PG-19, consists of Project Gutenberg books from before 1919 (Rae et al., 2019), which represent distinct styles from the more modern Books3 and BookCorpus. Additionally, the PG-19 dataset is already being used for long-distance context modeling.



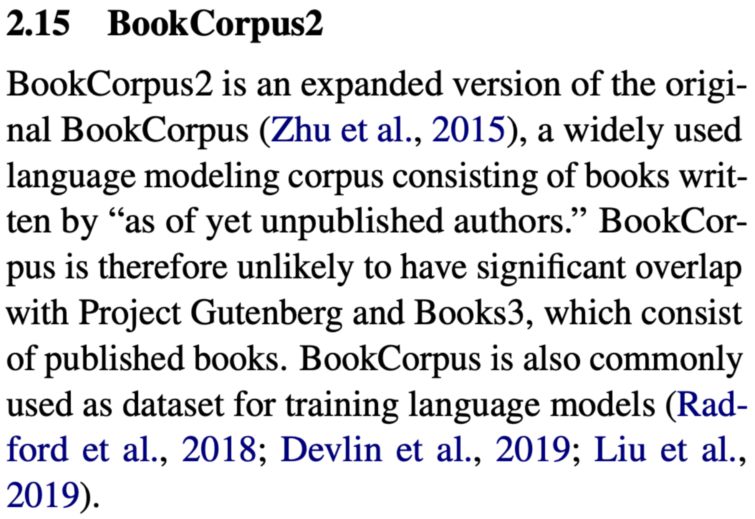
2.13 OpenSubtitles

The OpenSubtitles dataset is an English language dataset of subtitles from movies and television shows gathered by Tiedemann (2016). Subtitles provide an important source of natural dialog, as well as an understanding of fictional formats other than prose, which may prove useful for creative writing generation tasks such as screenwriting, speechwriting, and interactive storytelling.



2.14 DeepMind Mathematics

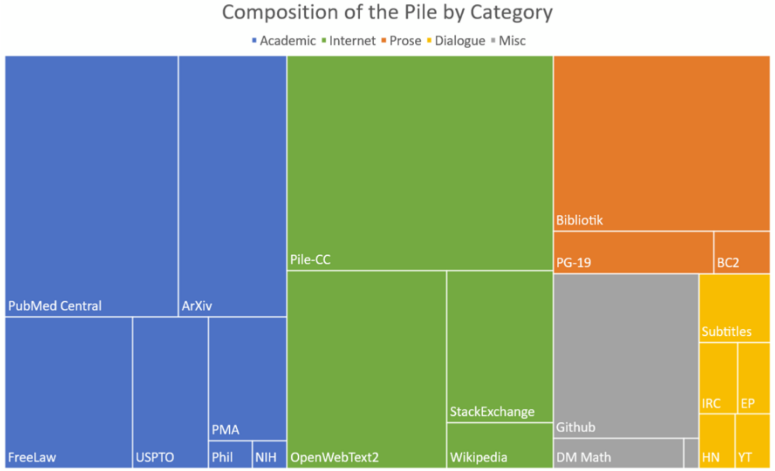
The DeepMind Mathematics dataset consists of a collection of mathematical problems from topics such as algebra, arithmetic, calculus, number theory, and probability, formatted as natural language prompts (Saxton et al., 2019). One major weakness of large language models has been performance on mathematical tasks (Brown et al., 2020), which may be due in part to a lack of math problems in the training set. By explicitly including a dataset of mathematical problems, we hope to improve the mathematical ability of language models trained on the Pile.

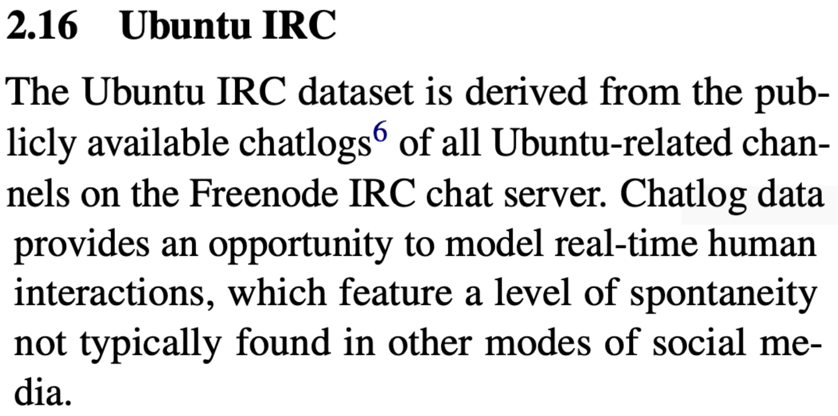


2.15 BookCorpus2

BookCorpus2 is an expanded version of the original BookCorpus (Zhu et al., 2015), a widely used language modeling corpus consisting of books written by "as of yet unpublished authors." BookCorpus is therefore unlikely to have significant overlap with Project Gutenberg and Books3, which consist of published books. BookCorpus is also commonly used as dataset for training language models (Radford et al., 2018; Devlin et al., 2019; Liu et al., 2019).

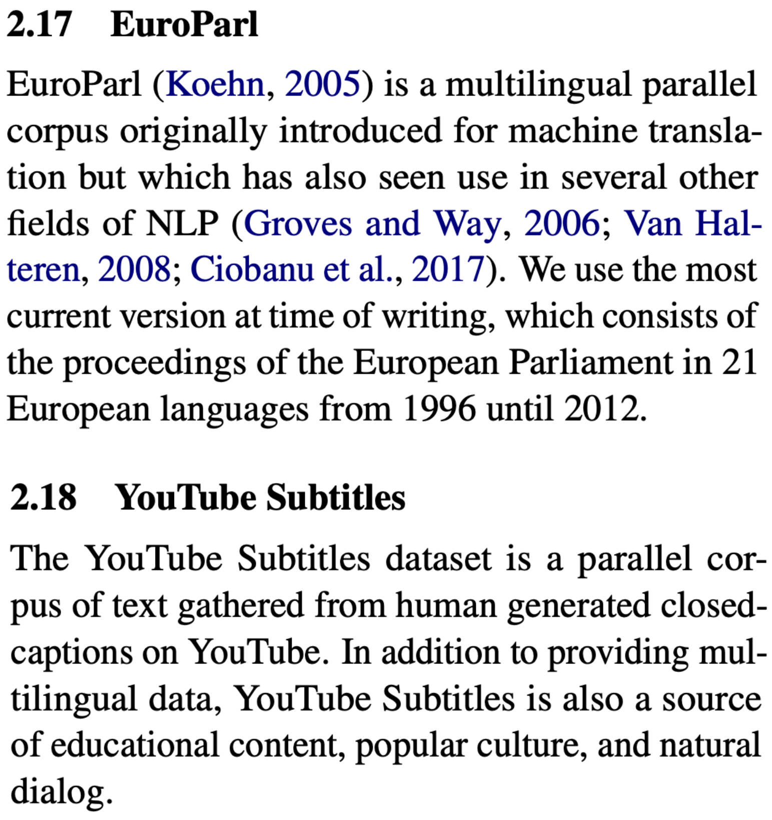
The Pile dataset - V





2.16 Ubuntu IRC

The Ubuntu IRC dataset is derived from the publicly available chatlogs[6] of all Ubuntu-related channels on the Freenode IRC chat server. Chatlog data provides an opportunity to model real-time human interactions, which feature a level of spontaneity not typically found in other modes of social media.

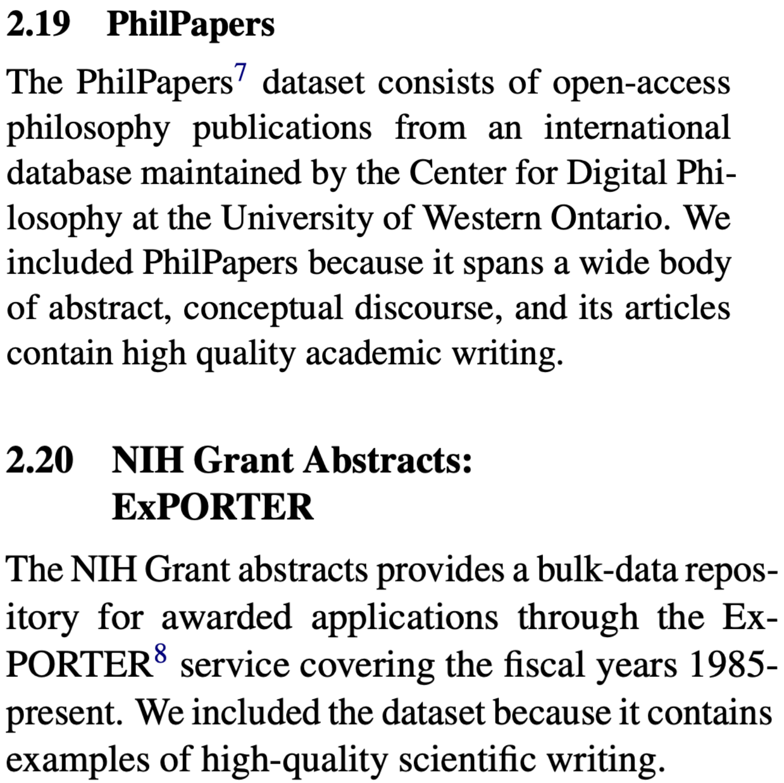


2.17 EuroParl

EuroParl (Koehn, 2005) is a multilingual parallel corpus originally introduced for machine translation but which has also seen use in several other fields of NLP (Groves and Way, 2006; Van Halteren, 2008; Ciobanu et al., 2017). We use the most current version at time of writing, which consists of the proceedings of the European Parliament in 21 European languages from 1996 until 2012.

2.18 YouTube Subtitles

The YouTube Subtitles dataset is a parallel corpus of text gathered from human generated closed-captions on YouTube. In addition to providing multilingual data, YouTube Subtitles is also a source of educational content, popular culture, and natural dialog.



2.19 PhilPapers

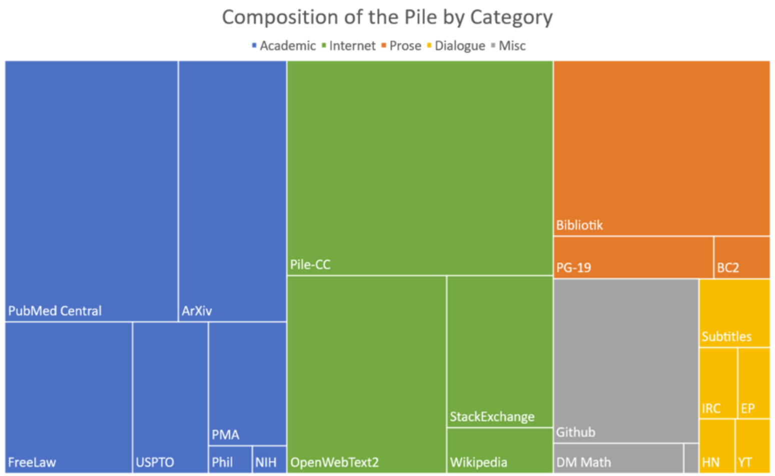
The PhilPapers[7] dataset consists of open-access philosophy publications from an international database maintained by the Center for Digital Philosophy at the University of Western Ontario. We included PhilPapers because it spans a wide body of abstract, conceptual discourse, and its articles contain high quality academic writing.

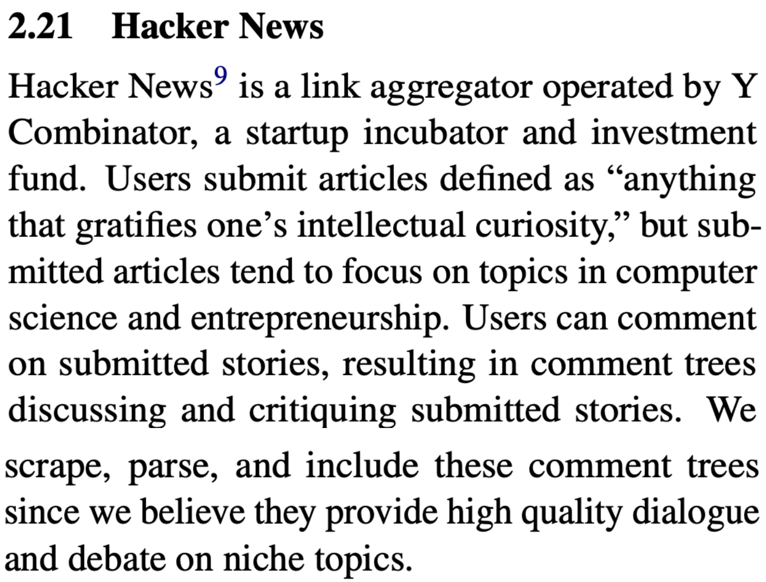
2.20 NIH Grant Abstracts: ExPORTER

The NIH Grant abstracts provides a bulk-data repository for awarded applications through the ExPORTER[8] service covering the fiscal years 1985-present. We included the dataset because it contains examples of high-quality scientific writing.

https://pile.eleuther.ai/paper.pdf

The Pile dataset - VI

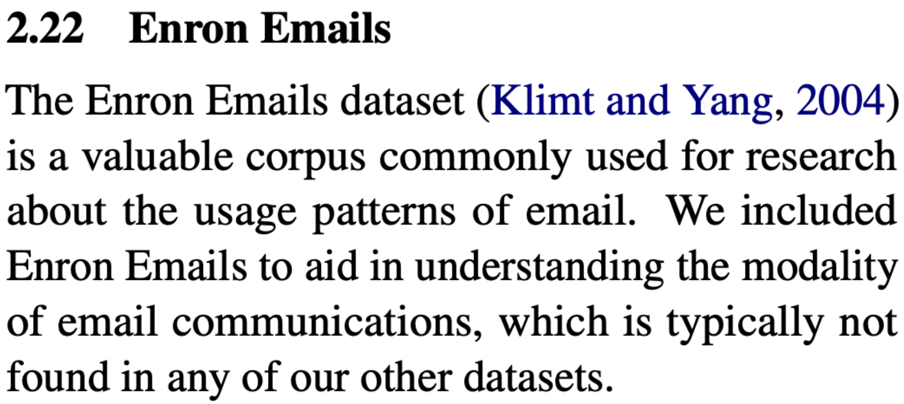




2.21 Hacker News

Hacker News[9] is a link aggregator operated by Y Combinator, a startup incubator and investment fund. Users submit articles defined as "anything that gratifies one's intellectual curiosity," but submitted articles tend to focus on topics in computer science and entrepreneurship. Users can comment on submitted stories, resulting in comment trees discussing and critiquing submitted stories. We scrape, parse, and include these comment trees since we believe they provide high quality dialogue and debate on niche topics.

https://pile.eleuther.ai/paper.pdf



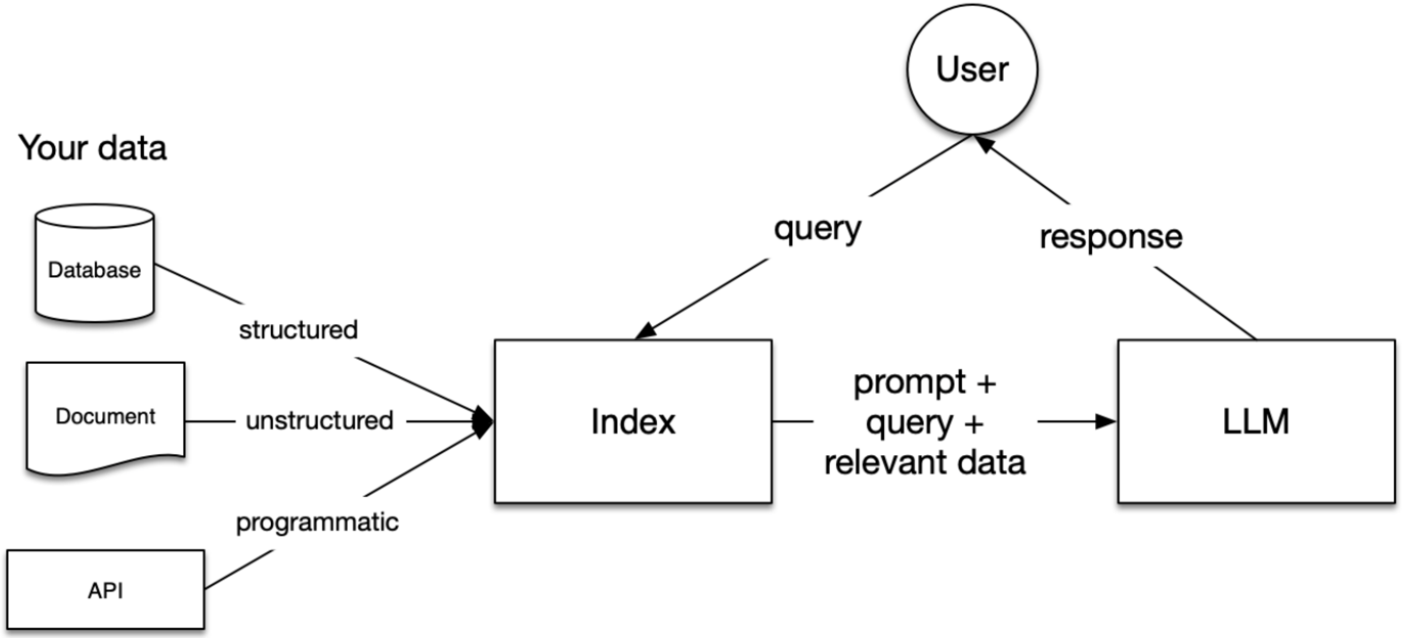
2.22 Enron Emails

The Enron Emails dataset (Klimt and Yang, 2004) is a valuable corpus commonly used for research about the usage patterns of email. We included Enron Emails to aid in understanding the modality of email communications, which is typically not found in any of our other datasets.

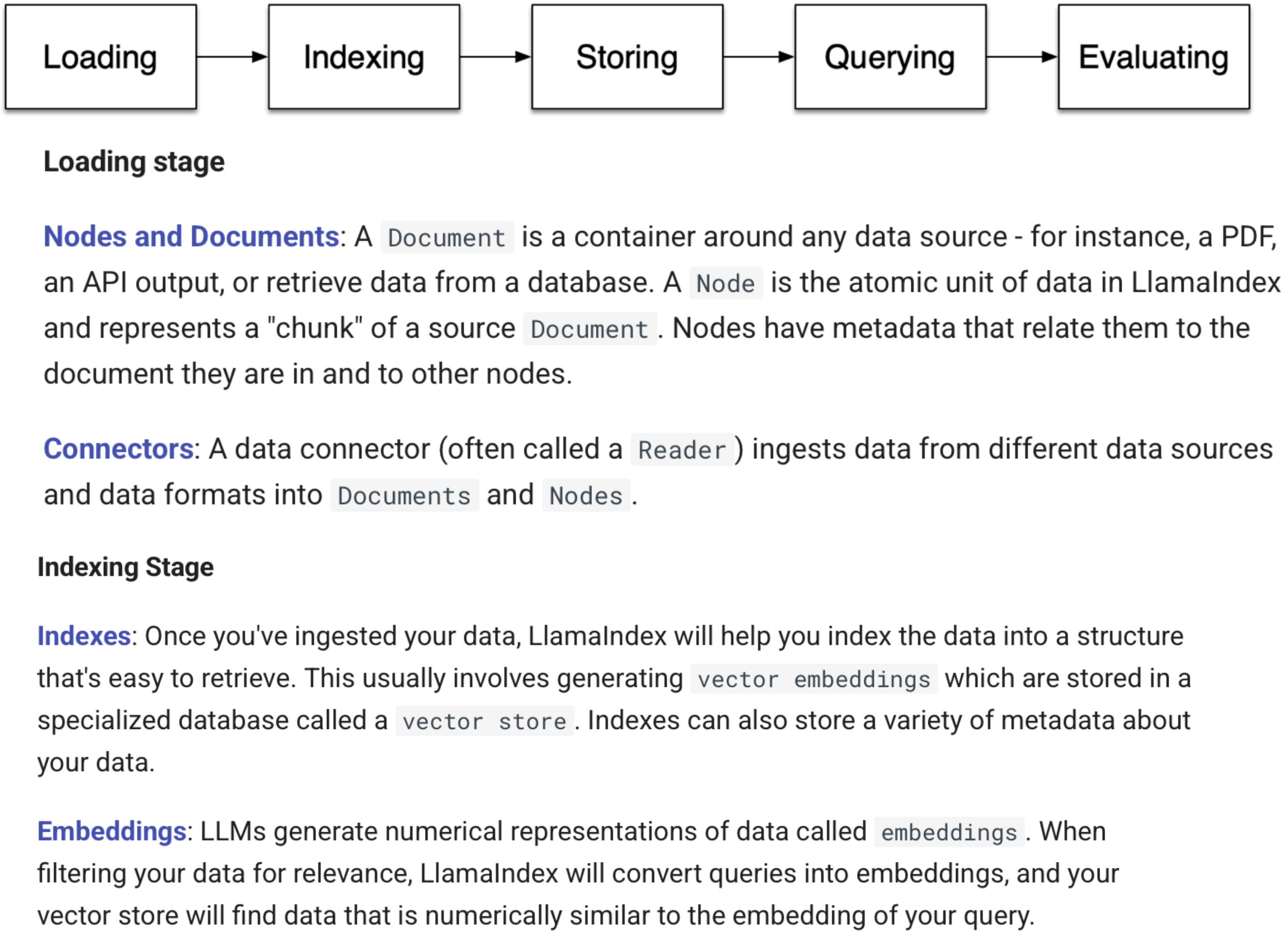
The Pile was originally developed to train EleutherAI's GPT-Neo models[8][9][10] but has become widely used to train other models, including Microsoft's Megatron-Turing Natural Language Generation,[11][12] Meta AI's Open Pre-trained Transformers,[13] LLaMA,[14] and Galactica,[15] Stanford University's BioMedLM 2.7B,[16] the Beijing Academy of Artificial Intelligence's Chinese-Transformer-XL,[17] Yandex's YaLM 100B,[18] and Apple's OpenELM.[19]

Llama + RAG

- LLMs are trained on enormous bodies of data but they aren't trained on **your** data. Retrieval-Augmented Generation (RAG) solves this problem by adding your data to the data LLMs already have access to. You will see references to RAG frequently in this documentation. Query engines, chat engines and agents often use RAG to complete their tasks.
- In RAG, your data is loaded and prepared for queries or "indexed". User queries act on the index, which filters your data down to the most relevant context. This context and your query then go to the LLM along with a prompt, and the LLM provides a response.



RAG – short intro I



Loading stage

Nodes and Documents: A `Document` is a container around any data source - for instance, a PDF, an API output, or retrieve data from a database. A `Node` is the atomic unit of data in LlamaIndex and represents a "chunk" of a source `Document`. Nodes have metadata that relate them to the document they are in and to other nodes.

Connectors: A data connector (often called a `Reader`) ingests data from different data sources and data formats into `Documents` and `Nodes`.

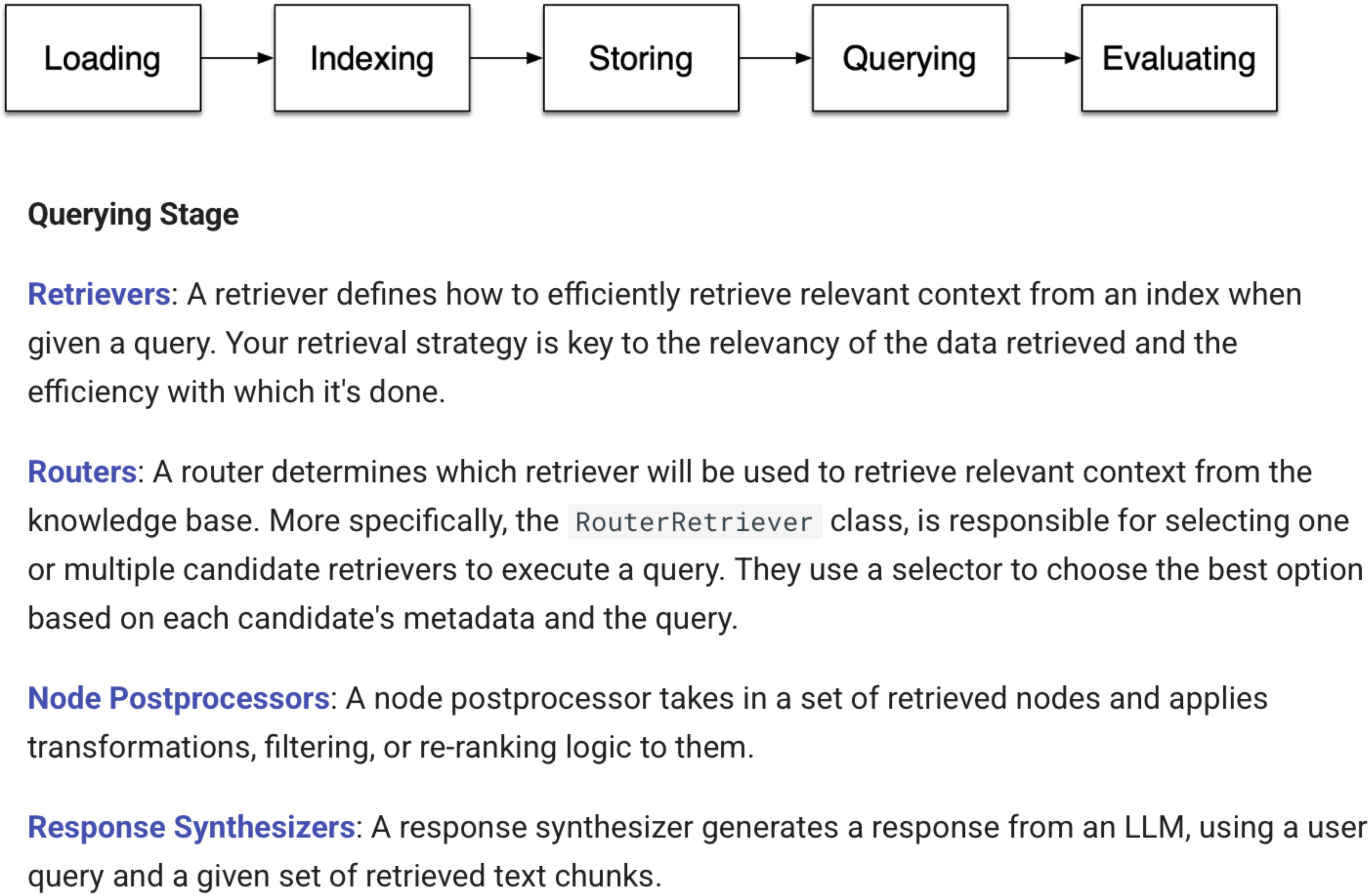
Indexing Stage

Indexes: Once you've ingested your data, LlamaIndex will help you index the data into a structure that's easy to retrieve. This usually involves generating `vector embeddings` which are stored in a specialized database called a `vector store`. Indexes can also store a variety of metadata about your data.

Embeddings: LLMs generate numerical representations of data called `embeddings`. When filtering your data for relevance, LlamaIndex will convert queries into embeddings, and your vector store will find data that is numerically similar to the embedding of your query.

https://docs.llamaindex.ai/en/stable/understanding/rag/

RAG – short intro II



Loading → Indexing → Storing → Querying → Evaluating

Querying Stage

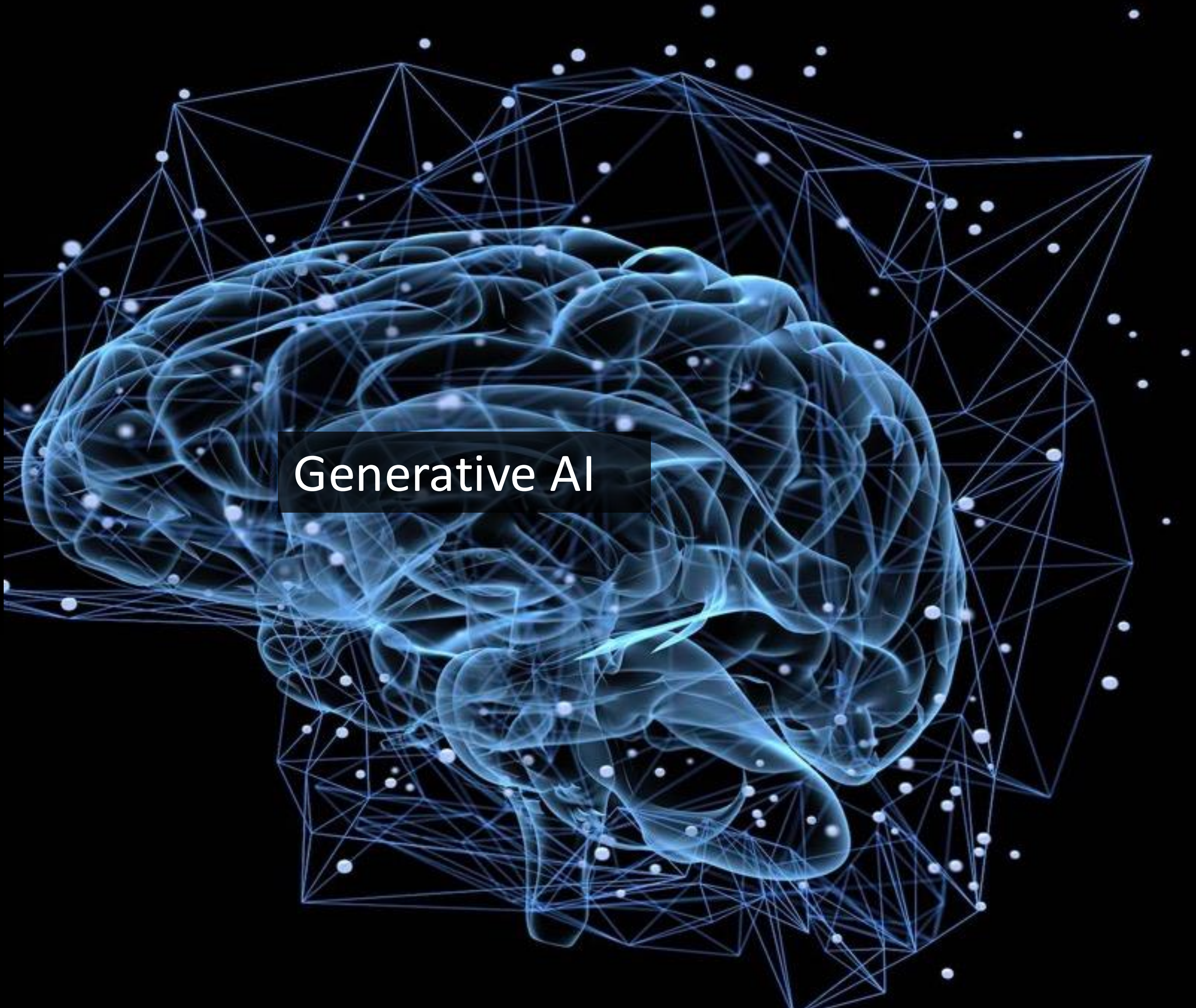
Retrievers: A retriever defines how to efficiently retrieve relevant context from an index when given a query. Your retrieval strategy is key to the relevancy of the data retrieved and the efficiency with which it's done.

Routers: A router determines which retriever will be used to retrieve relevant context from the knowledge base. More specifically, the `RouterRetriever` class, is responsible for selecting one or multiple candidate retrievers to execute a query. They use a selector to choose the best option based on each candidate's metadata and the query.

Node Postprocessors: A node postprocessor takes in a set of retrieved nodes and applies transformations, filtering, or re-ranking logic to them.

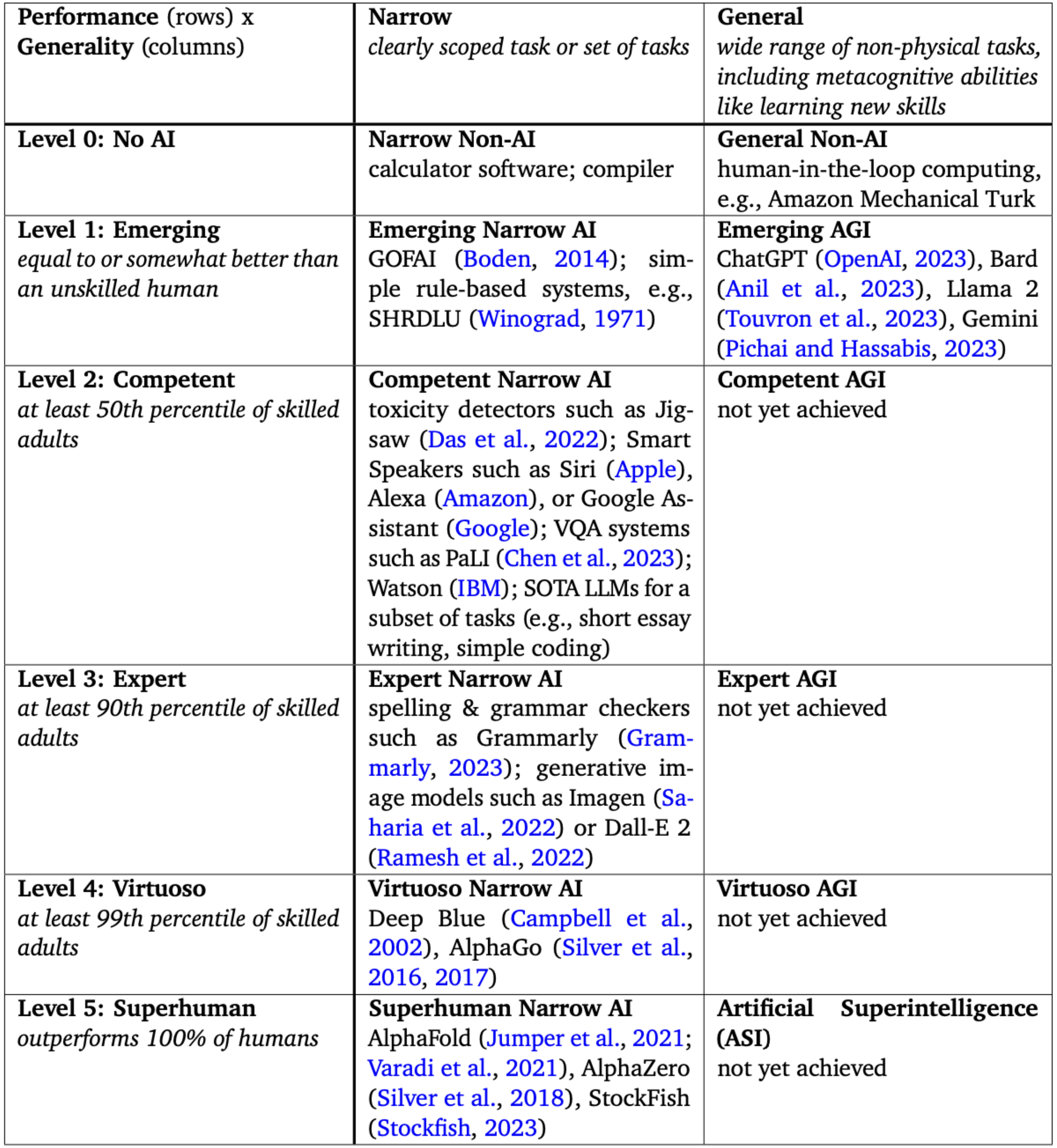
Response Synthesizers: A response synthesizer generates a response from an LLM, using a user query and a given set of retrieved text chunks.

https://docs.llamaindex.ai/en/stable/understanding/rag/



Generative AI

Levels of Artificial General Intelligence



Performance (rows) x **Generality** (columns)	**Narrow** *clearly scoped task or set of tasks*	**General** *wide range of non-physical tasks, including metacognitive abilities like learning new skills*
Level 0: No AI	**Narrow Non-AI** calculator software; compiler	**General Non-AI** human-in-the-loop computing, e.g., Amazon Mechanical Turk
Level 1: Emerging *equal to or somewhat better than an unskilled human*	**Emerging Narrow AI** GOFAI (Boden, 2014); simple rule-based systems, e.g., SHRDLU (Winograd, 1971)	**Emerging AGI** ChatGPT (OpenAI, 2023), Bard (Anil et al., 2023), Llama 2 (Touvron et al., 2023), Gemini (Pichai and Hassabis, 2023)
Level 2: Competent *at least 50th percentile of skilled adults*	**Competent Narrow AI** toxicity detectors such as Jigsaw (Das et al., 2022); Smart Speakers such as Siri (Apple), Alexa (Amazon), or Google Assistant (Google); VQA systems such as PaLI (Chen et al., 2023); Watson (IBM); SOTA LLMs for a subset of tasks (e.g., short essay writing, simple coding)	**Competent AGI** not yet achieved
Level 3: Expert *at least 90th percentile of skilled adults*	**Expert Narrow AI** spelling & grammar checkers such as Grammarly (Grammarly, 2023); generative image models such as Imagen (Saharia et al., 2022) or Dall-E 2 (Ramesh et al., 2022)	**Expert AGI** not yet achieved
Level 4: Virtuoso *at least 99th percentile of skilled adults*	**Virtuoso Narrow AI** Deep Blue (Campbell et al., 2002), AlphaGo (Silver et al., 2016, 2017)	**Virtuoso AGI** not yet achieved
Level 5: Superhuman *outperforms 100% of humans*	**Superhuman Narrow AI** AlphaFold (Jumper et al., 2021; Varadi et al., 2021), AlphaZero (Silver et al., 2018), StockFish (Stockfish, 2023)	**Artificial Superintelligence (ASI)** not yet achieved

Levels of AGI: https://arxiv.org/pdf/2311.02462.pdf

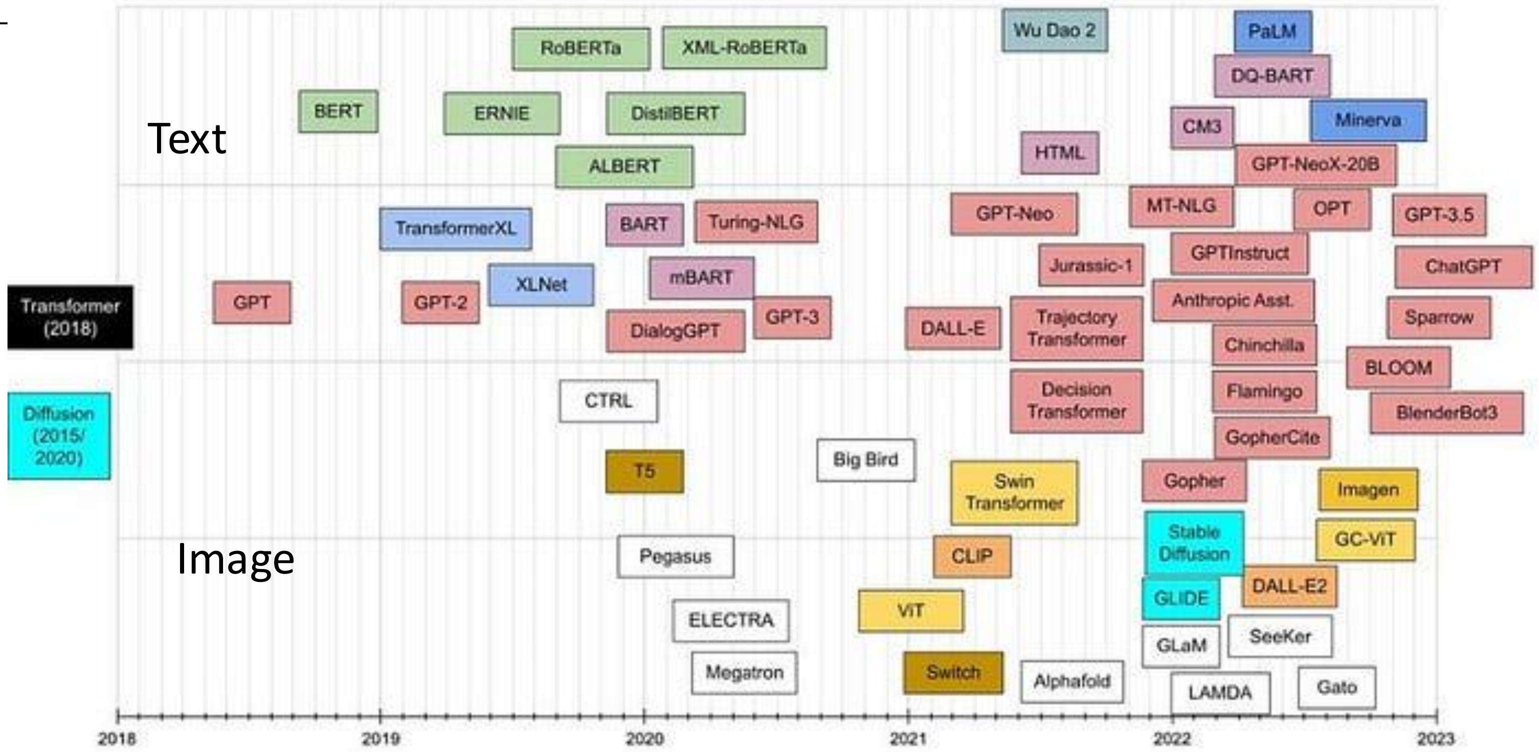
The Evolution of LLMs

1. In 2017, Google released the "Transformer Model", which can be used in question-answering systems, reading comprehension, sentiment analysis, instant translation of text or speech, and more

2. In 2018, OpenAI proposed "GPT" and Google proposed the "BERT" model, widely used in search engines, speech recognition, machine translation, question-answering systems, and more.

3. From 2018 to 2022, most of the research focused on BERT-related algorithms, when GPT performance was inferior to BERT

4. In 2023, ChatGPT (GPT3.5) was proposed by OpenAI, which significantly improves NLU's ability to understand most texts and surpasses humans in some area



In NLU

CNN

Local feature

RNN

Front and Back Dependency Issues

Self-Attention

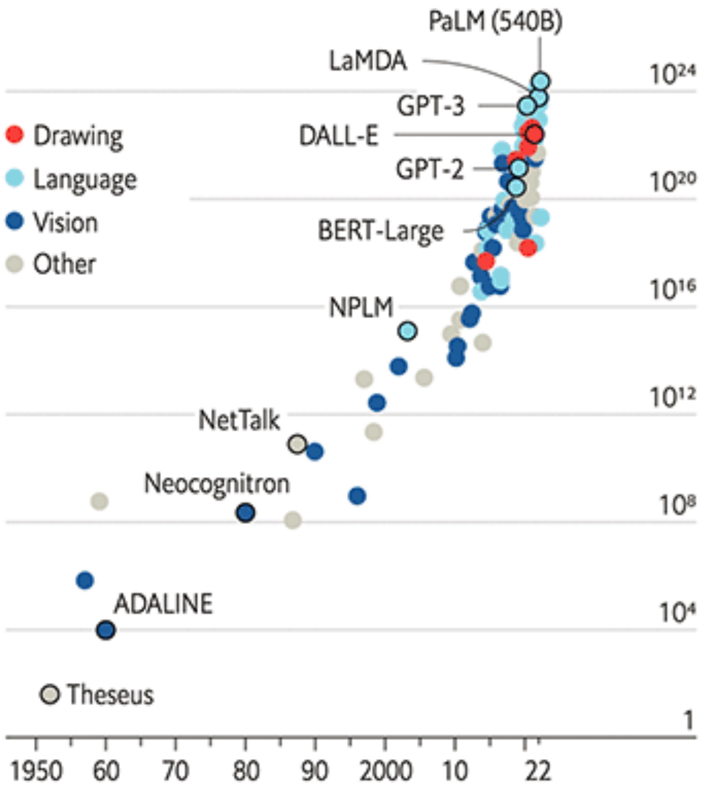
One to all attention, more flexible and trainable
need large datasets

The speed of development of Generative AI

The blessings of scale

AI training runs, estimated computing resources used

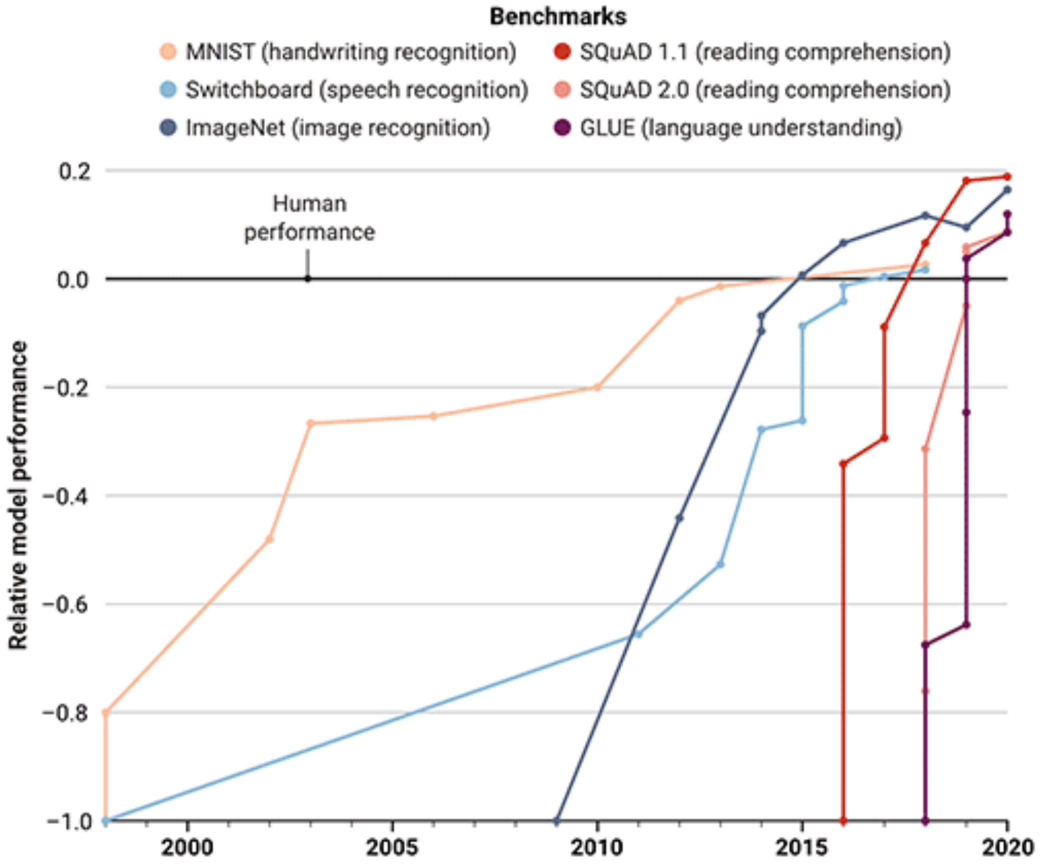
Floating-point operations, selected systems, by type, log scale



Sources: "Compute trends across three eras of machine learning", by J. Sevilla et al., arXiv, 2022; Our World in Data

Quick learners

The speed at which artificial intelligence models master benchmarks and surpass human baselines is accelerating. But they often fall short in the real world.



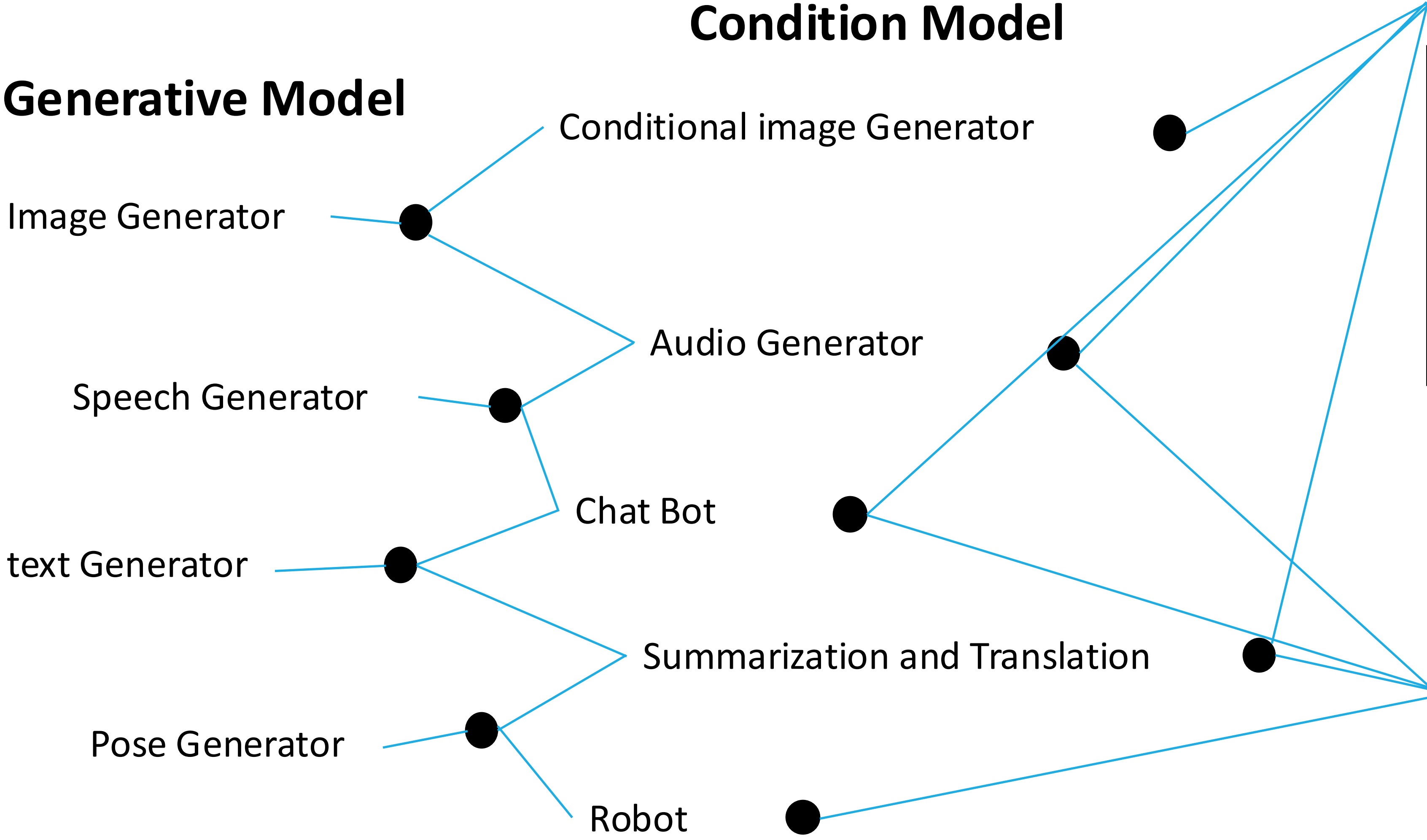
(GRAPHIC) K. FRANKLIN/*SCIENCE*; (DATA) D. KIELA *ET AL.*, DYNABENCH: RETHINKING BENCHMARKING IN NLP, DOI:10.48550/ARXIV.2104.14337

Generative AI Application

Multi-Model

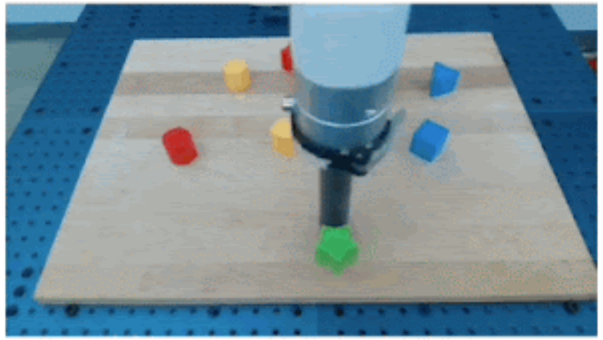
Condition Model

Generative Model



NLU + Image Generator

NLU + Robot

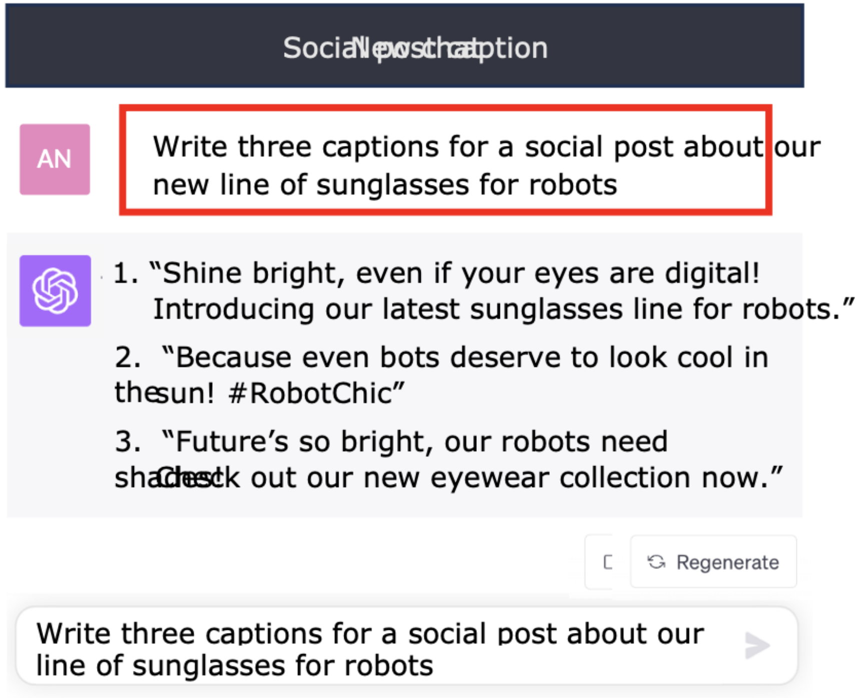


push the green star to the bottom center

What is Generative AI

Artificial intelligence systems that can produce high quality content, specifically **text, images, and audio.**

Prompts



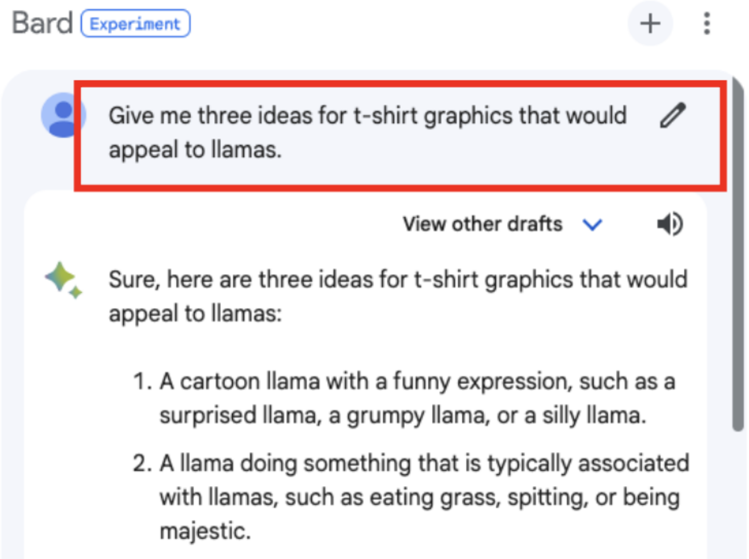
Social post caption / New chat

AN Write three captions for a social post about our new line of sunglasses for robots

1. “Shine bright, even if your eyes are digital! Introducing our latest sunglasses line for robots.”
2. “Because even bots deserve to look cool in the sun! #RobotChic”
3. “Future’s so bright, our robots need shades! Check out our new eyewear collection now.”

Write three captions for a social post about our line of sunglasses for robots

ChatGPT/OpenAI

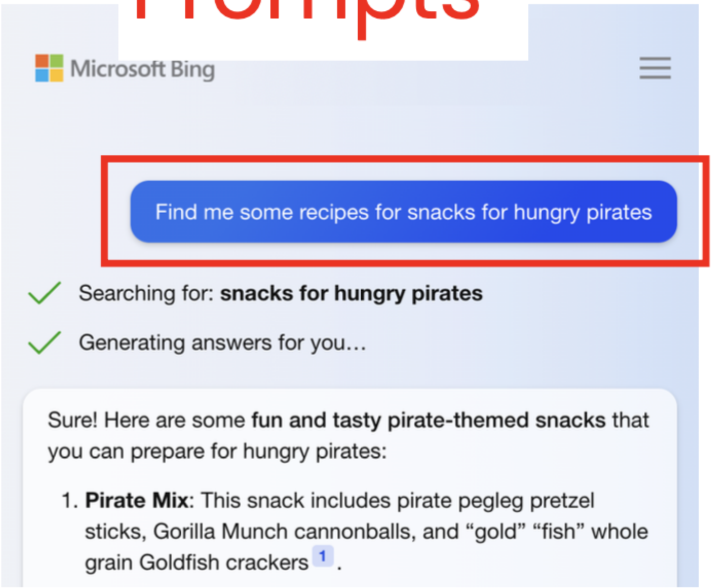


Give me three ideas for t-shirt graphics that would appeal to llamas.

Sure, here are three ideas for t-shirt graphics that would appeal to llamas:

1. A cartoon llama with a funny expression, such as a surprised llama, a grumpy llama, or a silly llama.
2. A llama doing something that is typically associated with llamas, such as eating grass, spitting, or being majestic.

Bard/Google



Find me some recipes for snacks for hungry pirates

✓ Searching for: **snacks for hungry pirates**

✓ Generating answers for you…

Sure! Here are some **fun and tasty pirate-themed snacks** that you can prepare for hungry pirates:

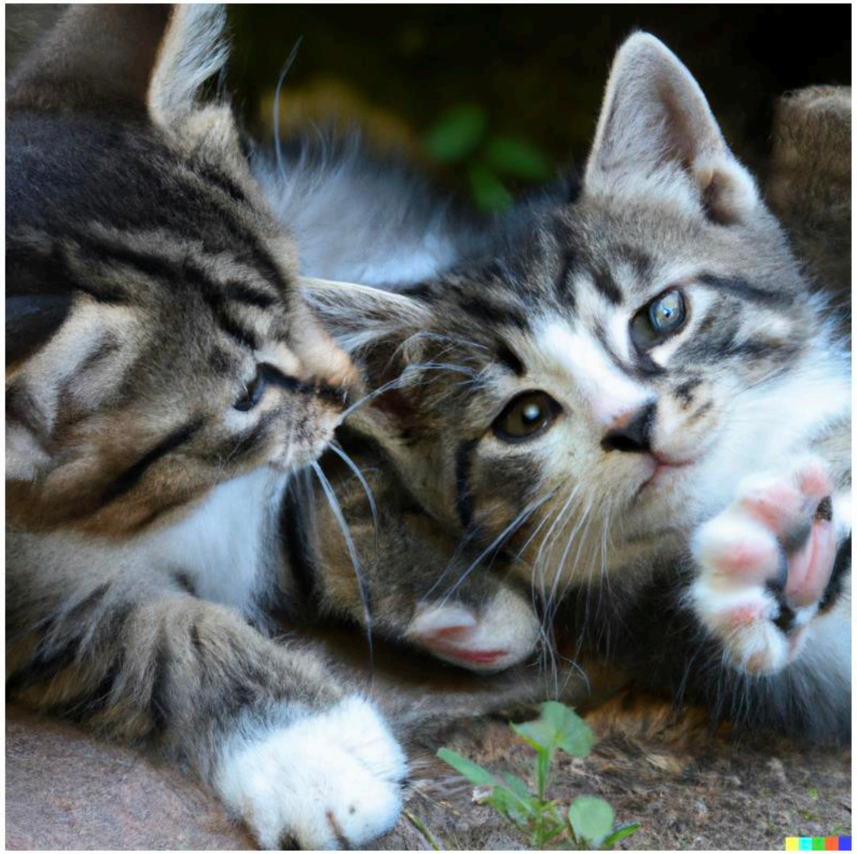
1. **Pirate Mix**: This snack includes pirate pegleg pretzel sticks, Gorilla Munch cannonballs, and “gold” “fish” whole grain Goldfish crackers [1].

Bing Chat/Microsoft

Multimedia Generation

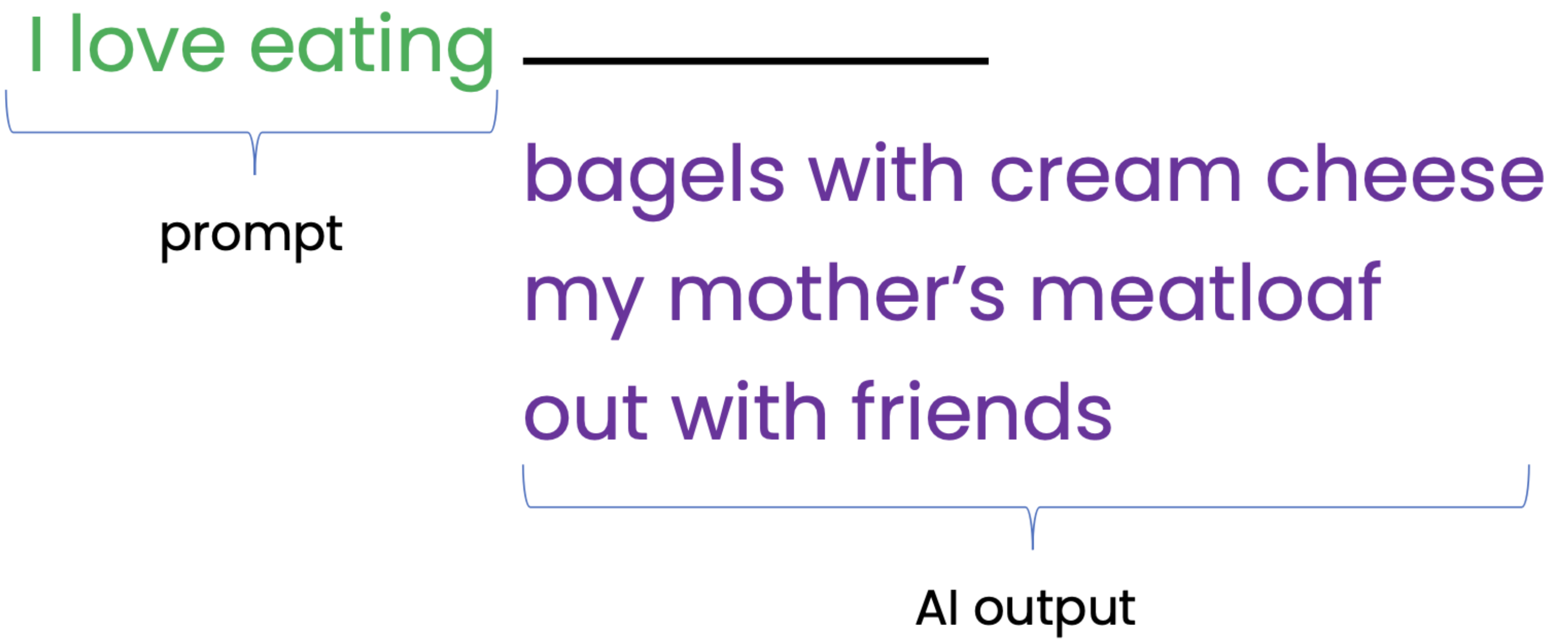
A beautiful, pastoral mountain scene. Landscape painting style (Midjourney)

Two cute kittens playing (DALL-E)



Generating Text using Large Language Models

Text generation process



I love eating ____________

prompt

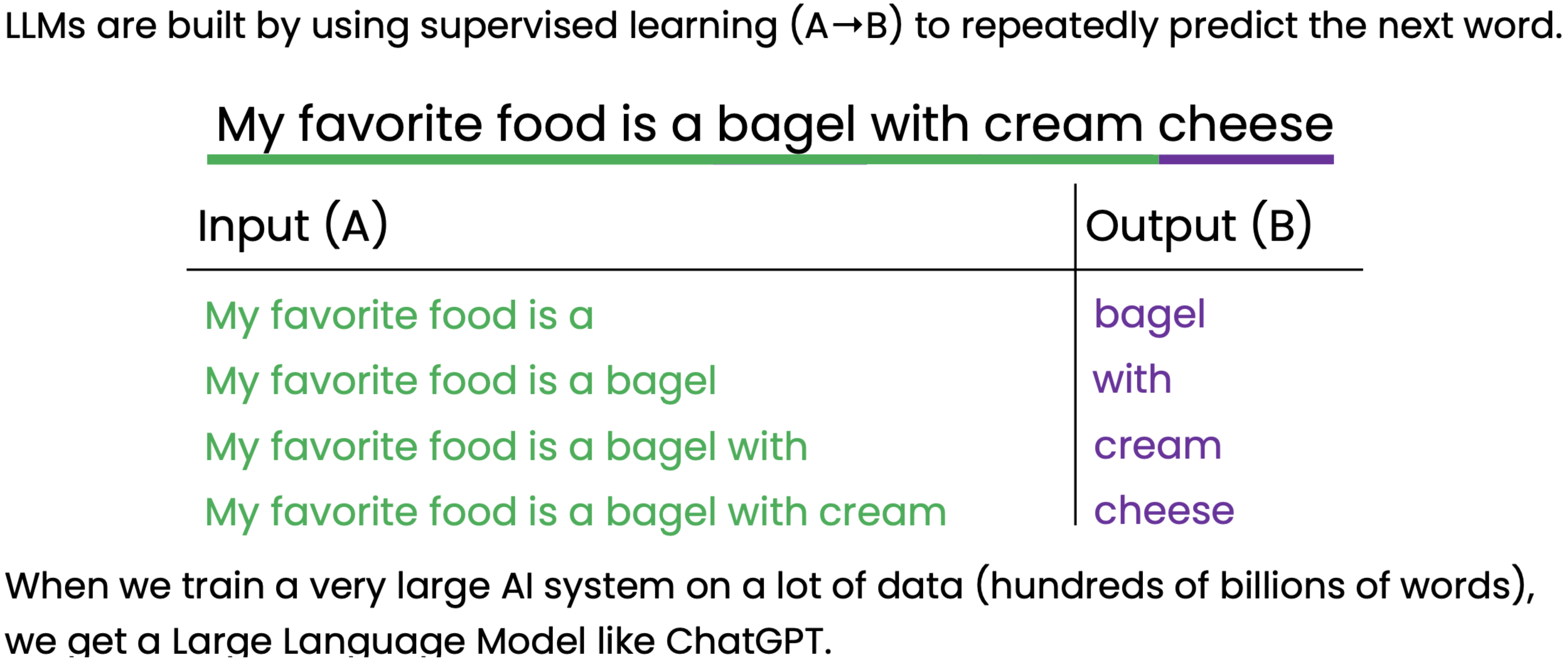
bagels with cream cheese

my mother’s meatloaf

out with friends

AI output

How LLM works



LLMs are built by using supervised learning (A→B) to repeatedly predict the next word.

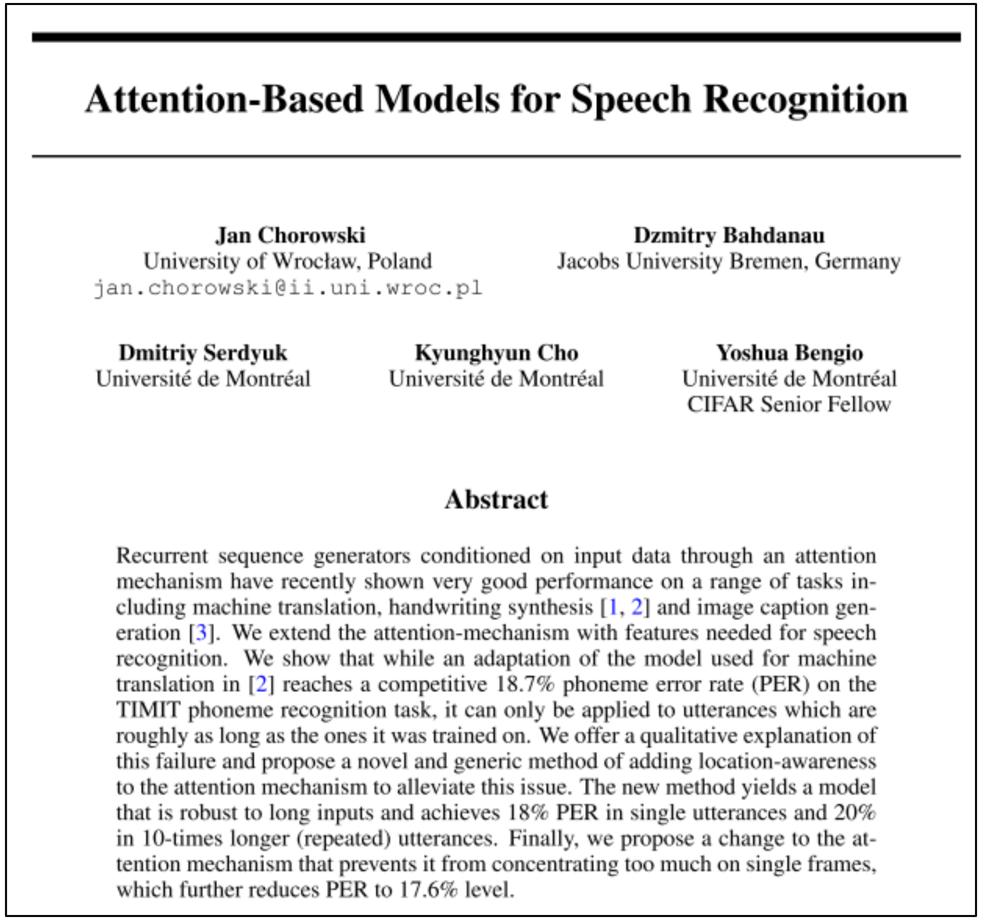
My favorite food is a bagel with cream cheese

Input (A)	Output (B)
My favorite food is a	bagel
My favorite food is a bagel	with
My favorite food is a bagel with	cream
My favorite food is a bagel with cream	cheese

When we train a very large AI system on a lot of data (hundreds of billions of words), we get a Large Language Model like ChatGPT.

Attention Model

[Bengio_2015]



Attention-Based Models for Speech Recognition

Jan Chorowski
University of Wrocław, Poland
jan.chorowski@ii.uni.wroc.pl

Dzmitry Bahdanau
Jacobs University Bremen, Germany

Dmitriy Serdyuk
Université de Montréal

Kyunghyun Cho
Université de Montréal

Yoshua Bengio
Université de Montréal
CIFAR Senior Fellow

Abstract

Recurrent sequence generators conditioned on input data through an attention mechanism have recently shown very good performance on a range of tasks including machine translation, handwriting synthesis [1, 2] and image caption generation [3]. We extend the attention-mechanism with features needed for speech recognition. We show that while an adaptation of the model used for machine translation in [2] reaches a competitive 18.7% phoneme error rate (PER) on the TIMIT phoneme recognition task, it can only be applied to utterances which are roughly as long as the ones it was trained on. We offer a qualitative explanation of this failure and propose a novel and generic method of adding location-awareness to the attention mechanism to alleviate this issue. The new method yields a model that is robust to long inputs and achieves 18% PER in single utterances and 20% in 10-times longer (repeated) utterances. Finally, we propose a change to the attention mechanism that prevents it from concentrating too much on single frames, which further reduces PER to 17.6% level.

In 2015, Bengio 's Model focuses on every phenon's recognition as the combined weights.

$$\boxed{\begin{aligned} \alpha_i &= Attend(s_{i-1}, \alpha_{i-1}, h) \\ g_i &= \sum_{j=1}^{L} \alpha_{i,j} h_j \\ y_i &\sim Generate(s_{i-1}, g_i), \end{aligned}}$$

h : Input

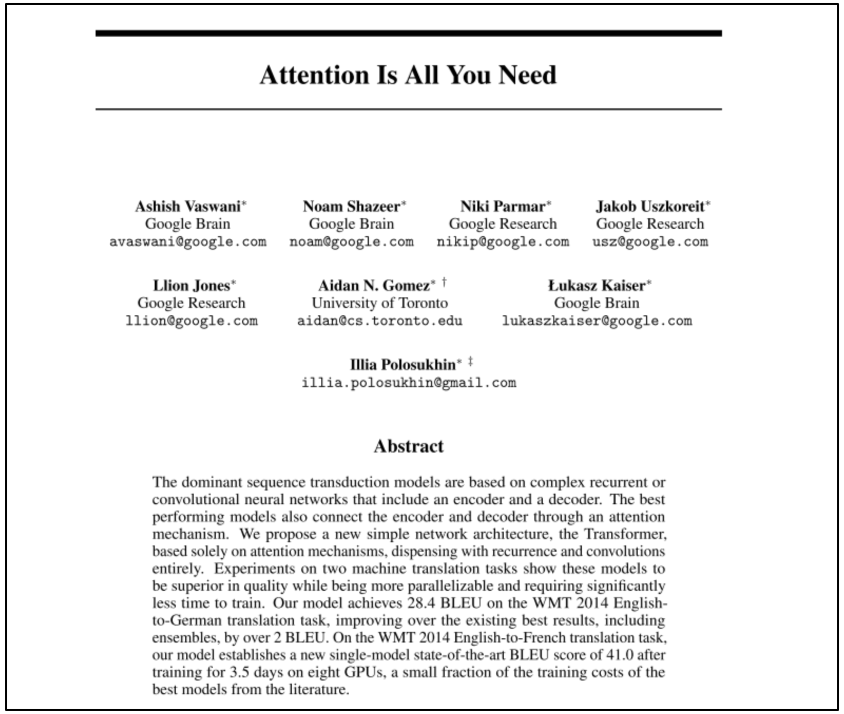
α_i: Attention Weight

y_i: Output

Chorowski, Jan K., et al. "Attention-based models for speech recognition." *Advances in neural information processing systems* 28 (2015).

Transformer [Vaswani_2017]

In 2017, 8 Google researchers proposed Transformer Neuron Networks based on Attention, which was adopted by ChatGPT.



Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

Cited 66157 (2023/2/21)

Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Jakob Uszkoreit proposed replacing RNNs with **self-attention** and started the effort to evaluate this idea.

Noam Shazeer proposed **scaled dot-product attention**, **multi-head attention** and the **parameter-free position representation**.

Transformer

- Transformer is a Deep Learning Model based on Self-Attention
- **Transformer** encodes and decodes data with different weights.
- Examples of **transformer language models include: GPT (GPT-1、 GPT-2、 GPT-3、 ChatGPT) and BERT models (BERT、 RoBERTa 、 ERNIE).**

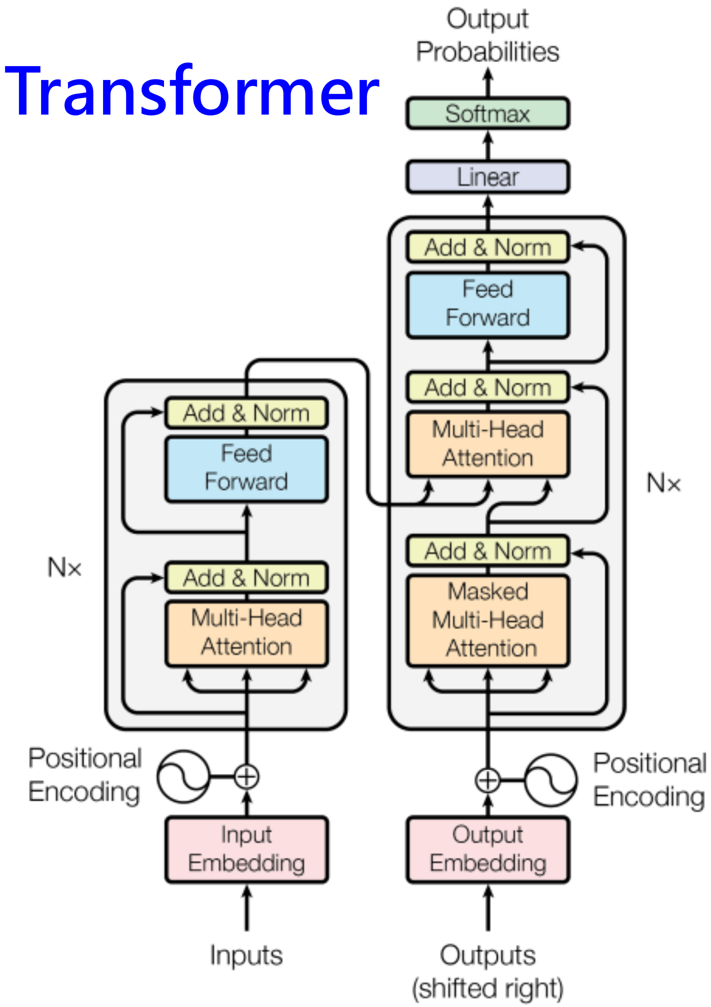
Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Transformer

Encoder

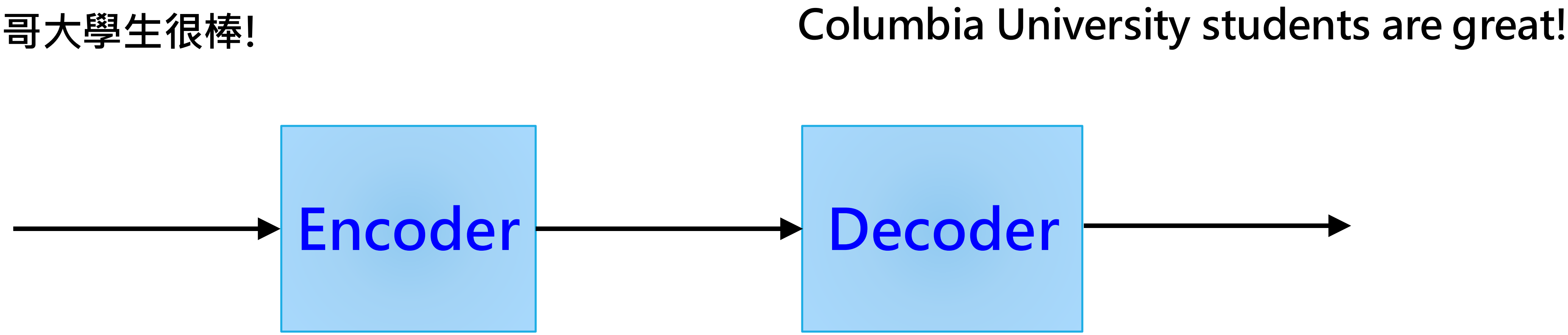
Transformer

Decoder



Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Transformer



Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Transformer

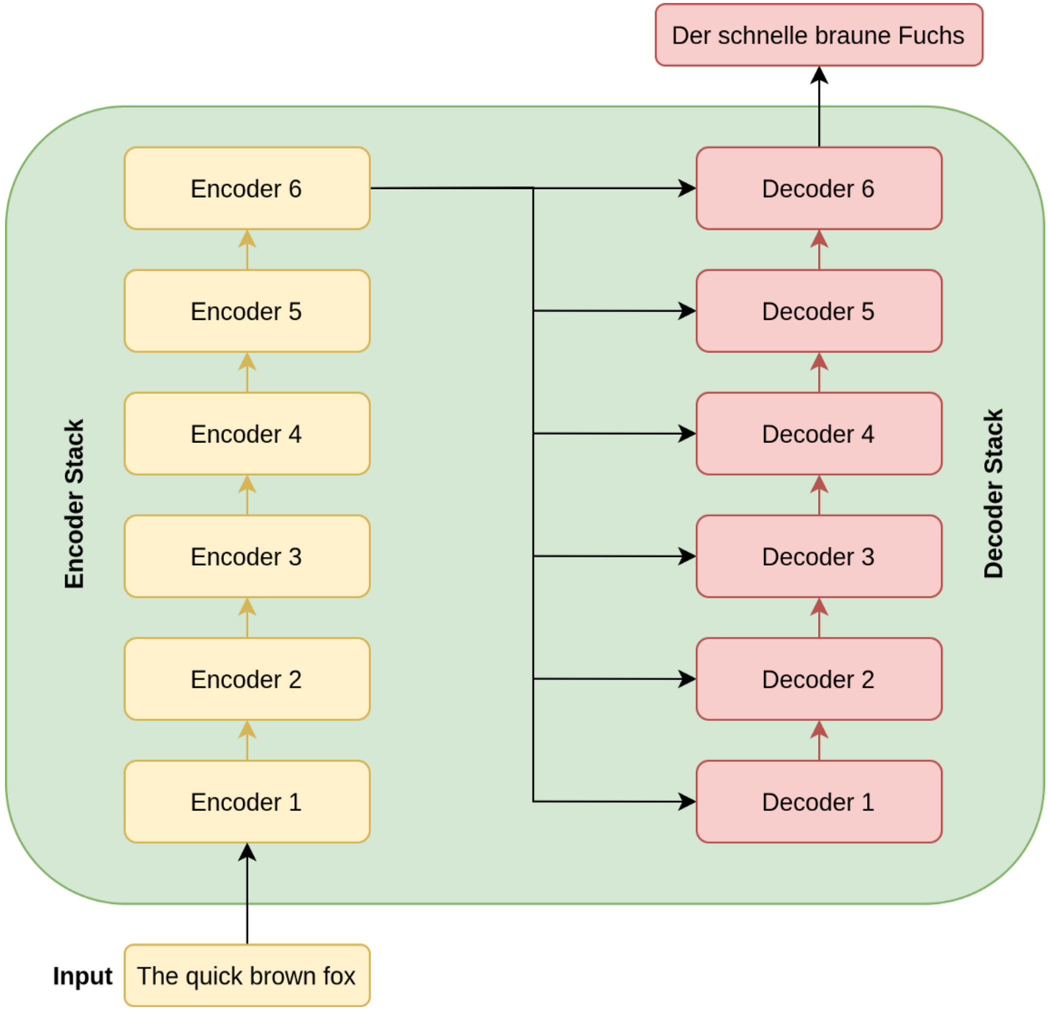
Attention

K

Q

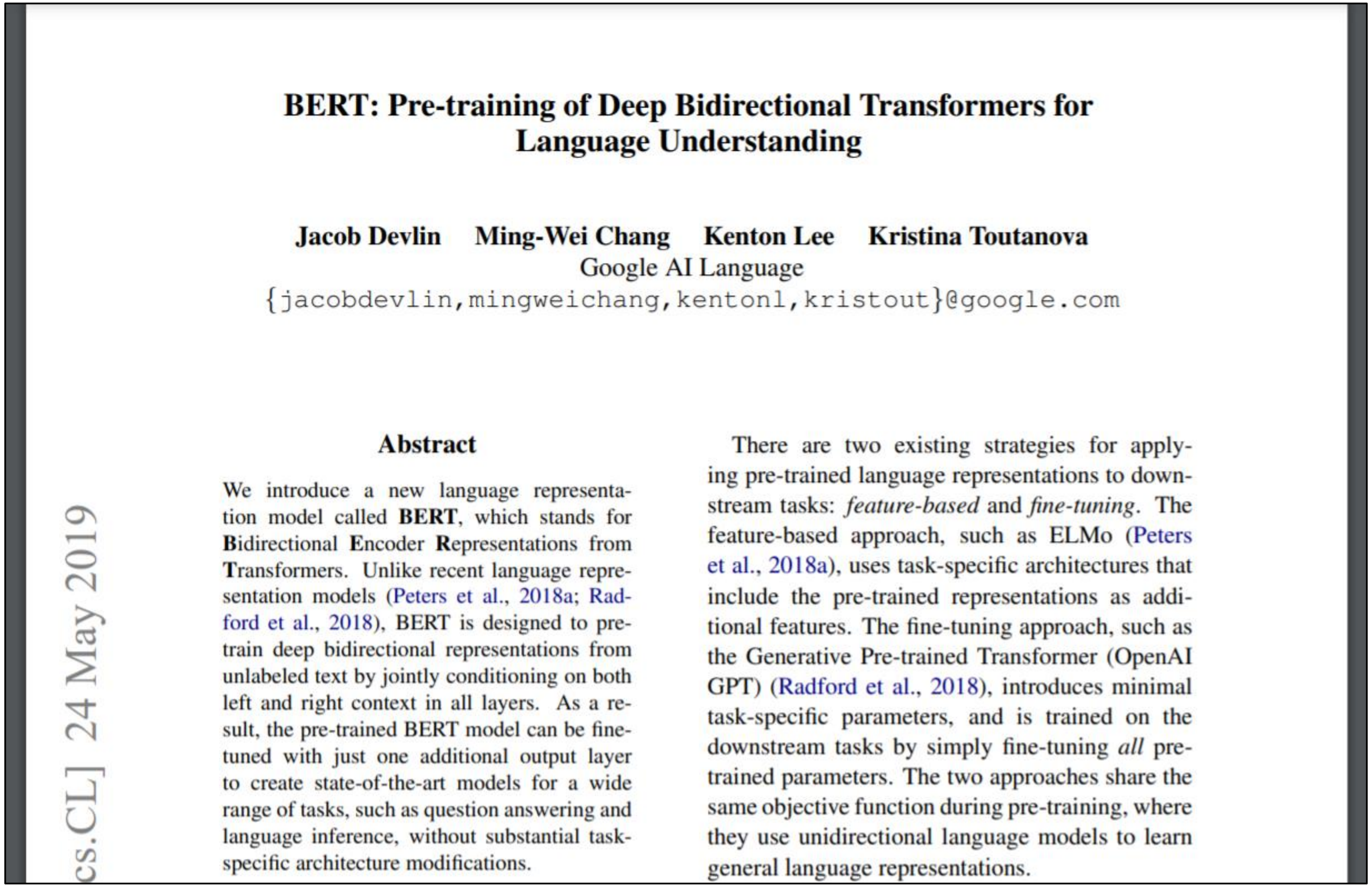
		k_1	k_2	k_3	k_4	k_5	k_6
	weights	Columbia	university	students	are	great	!
q_1	哥	1	0.5	0.2	0	0.3	0.2
q_2	大	0.5	1	0.2	0.1	0.3	0.1
q_3	學	0.2	0.2	1	0	0.5	0.2
q_4	生	0.3	0.3	0.8	0.5	0.5	0.6
q_5	很	0	0.1	0	1	0.5	0
q_6	棒	0.3	0.3	0.5	0.5	1	0.8
q_7	!	0.2	0.1	0.2	0	0.8	1

Transformer Translation



Transformer uses 6 layers of encoder and decoder to achieve the same quality of SOTA English-German and English-French translation.

BERT Introduction



BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Abstract

We introduce a new language representation model called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning *all* pre-trained parameters. The two approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

BERT Introduction

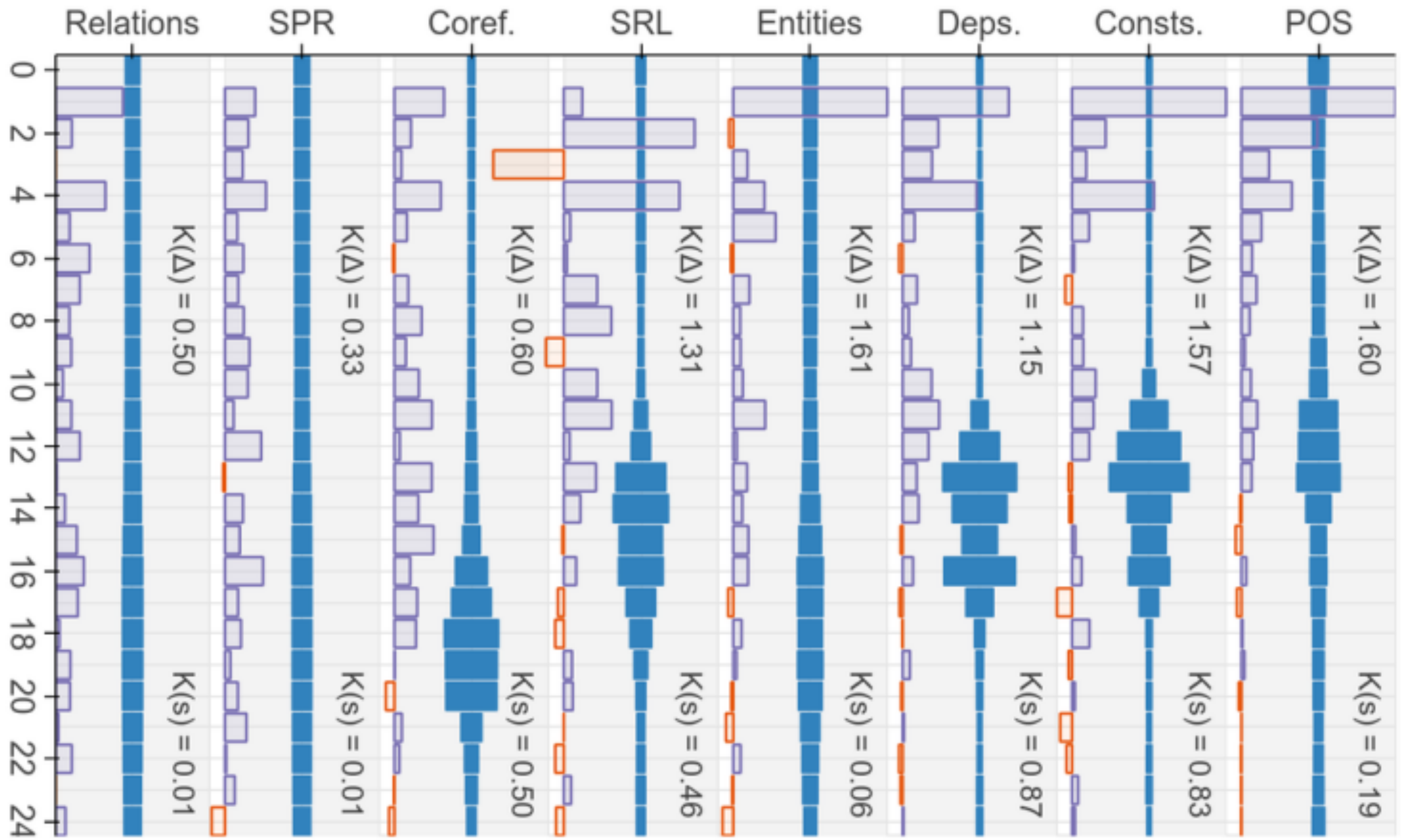
- 2018 Google'BERT has 24 layers of Transformer Encoder
- BERT's original model is based on Wikipedia and booksorpus, using unsupervised training to create BERT.
- At Stanford's Machine Reasoning Test SQuAD1.1 beats human performance.
- Google NLU English was replaced from seq2seq to BERT

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

BERT understands language's meaning

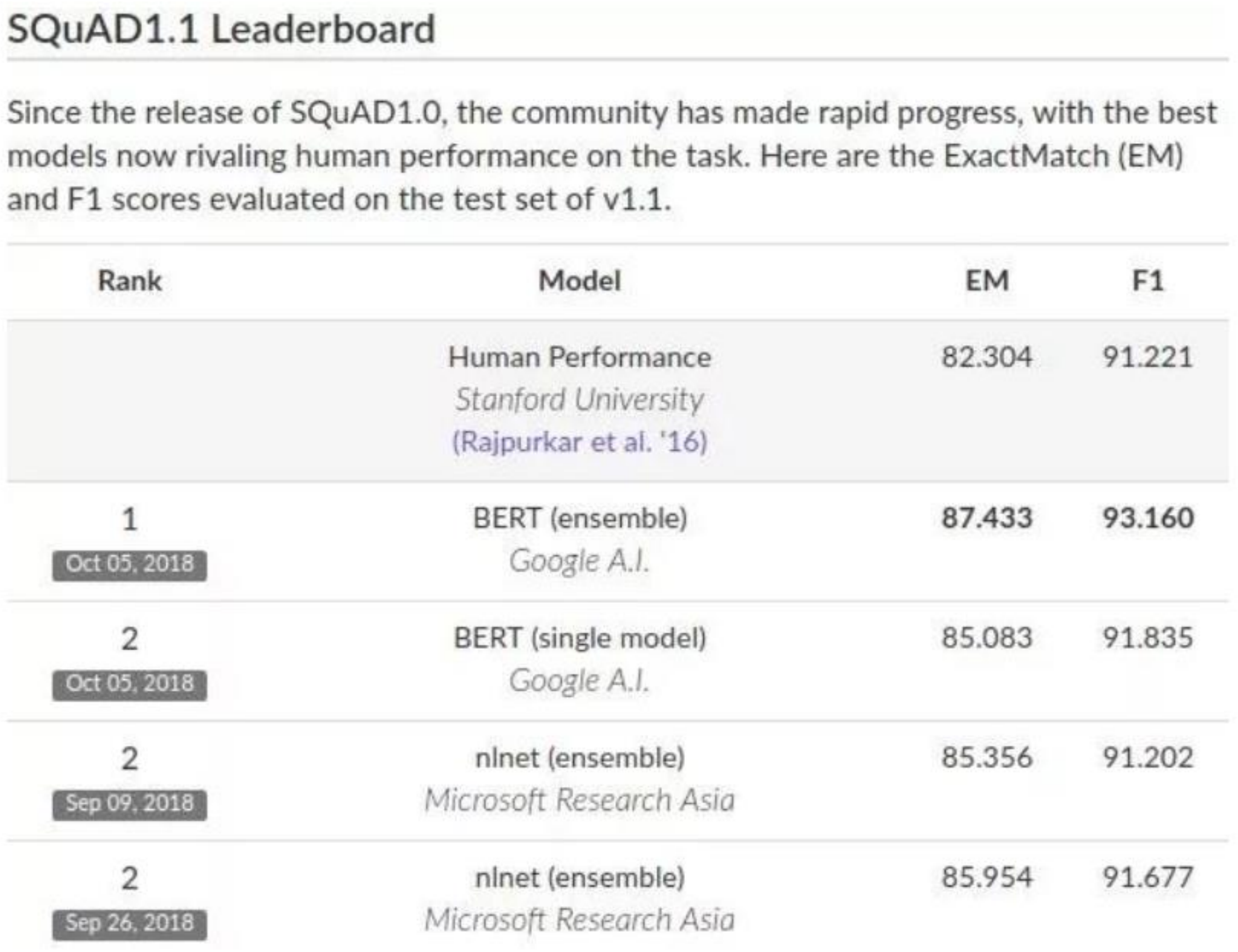
High-Level NLP ⟷ Low-Level NLP

Semantic-Role Core Semantic-Role Level Dependents Constitutions



Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. *arXiv preprint arXiv:1905.05950*.

In 2018, BERT Comprehension test outperformed human



SQuAD1.1 Leaderboard

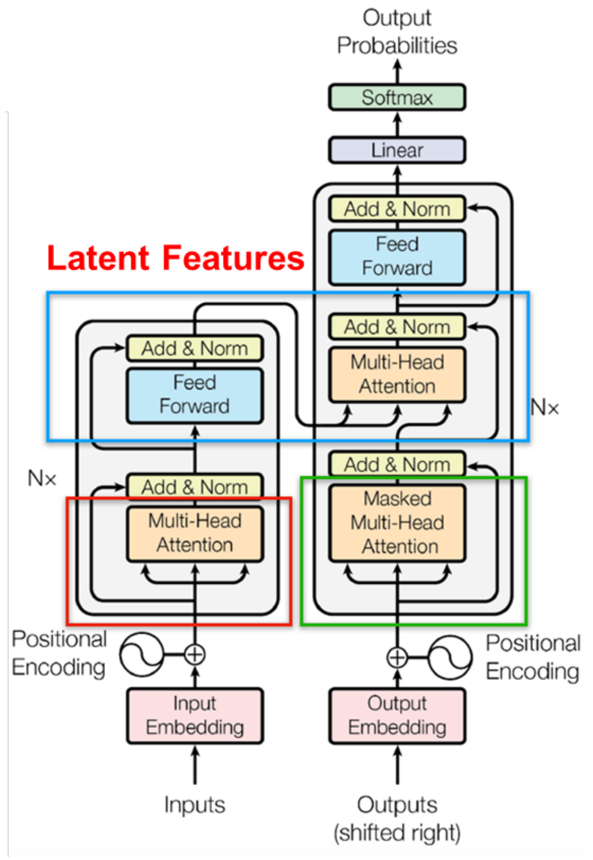
Since the release of SQuAD1.0, the community has made rapid progress, with the best models now rivaling human performance on the task. Here are the ExactMatch (EM) and F1 scores evaluated on the test set of v1.1.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google A.I.*	**87.433**	**93.160**
2 Oct 05, 2018	BERT (single model) *Google A.I.*	85.083	91.835
2 Sep 09, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.356	91.202
2 Sep 26, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.954	91.677

Transformer to GPT

Transformer

Input -> **Encoder** -> Latent Feature + Masked Output -> **Decoder** -> Output



An [redacted] a day keeps the doctor away

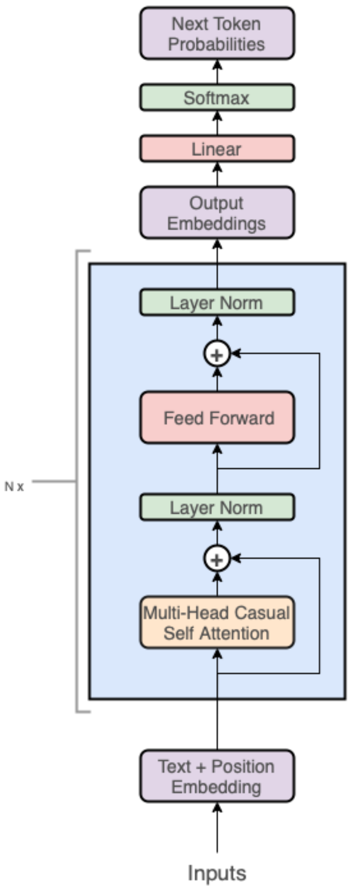
apple 95%
banana 5%

Masked Language Learning

An apple a day keeps the doctor away

GPT

Input -> **Decoder(with Casual mask)** -> shift Output



An

apple 99%
almond 1%

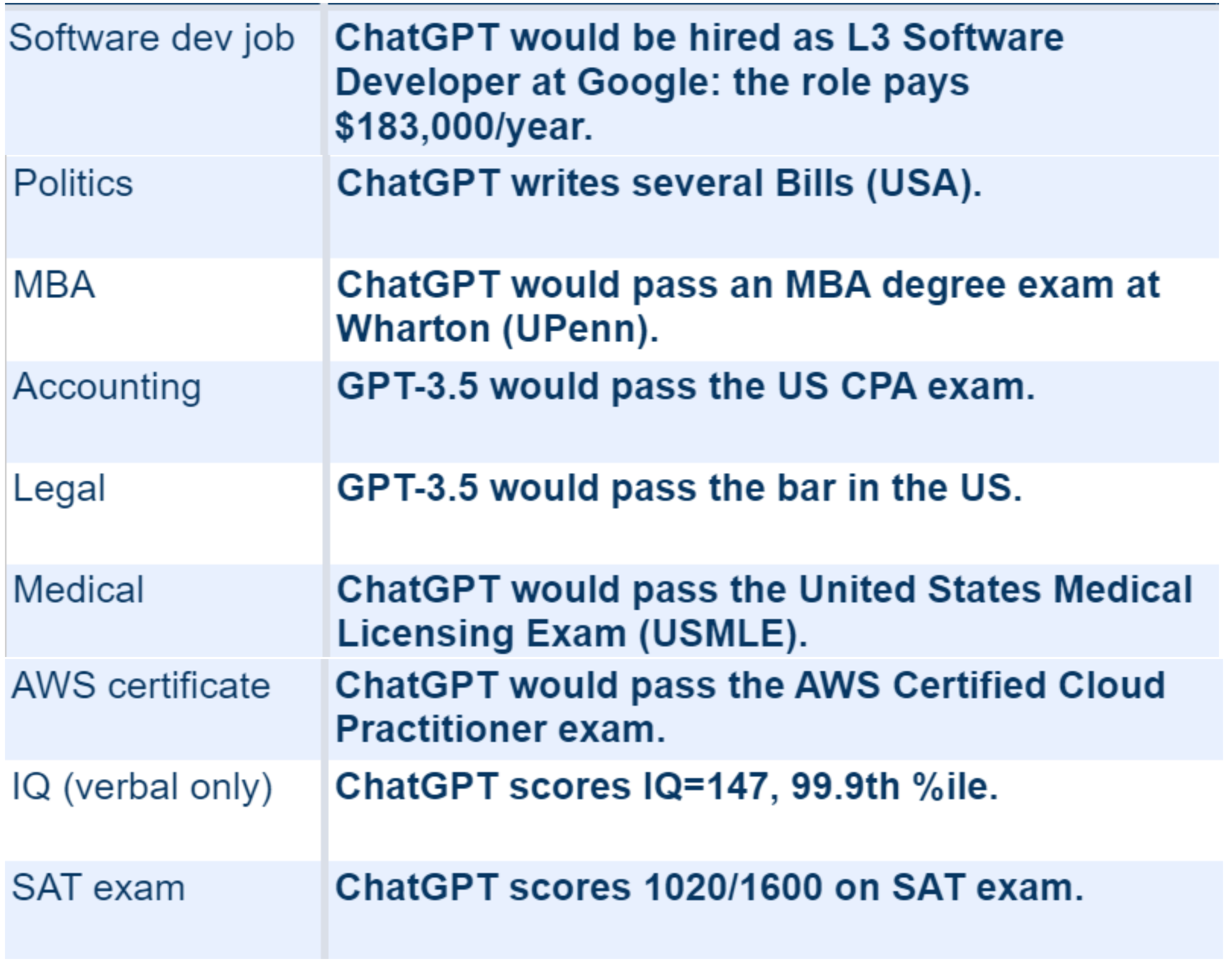
An apple

a 99%
watch 1%

An apple a

Autoregressive Learning

ChatGPT

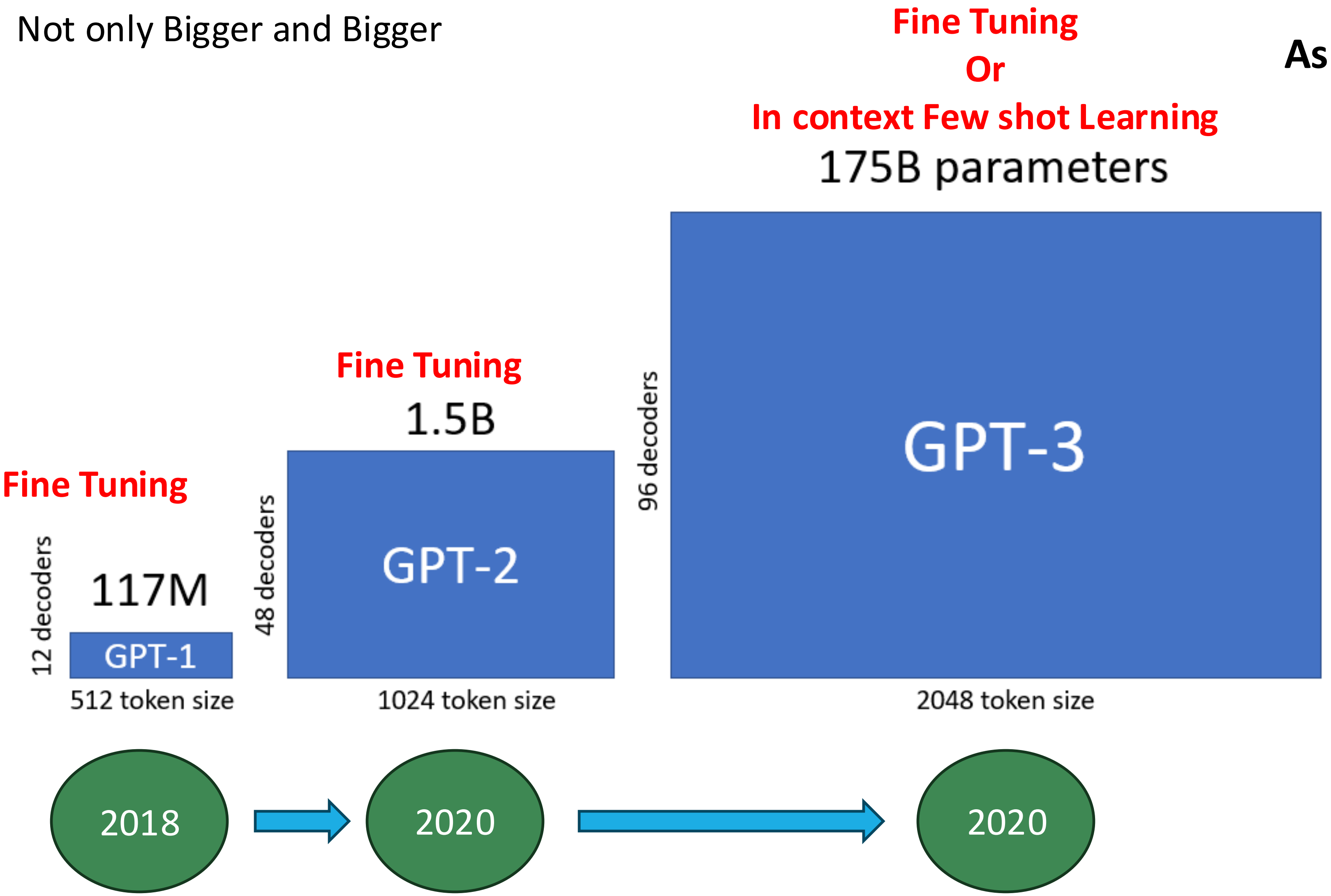


Software dev job	**ChatGPT would be hired as L3 Software Developer at Google: the role pays $183,000/year.**
Politics	**ChatGPT writes several Bills (USA).**
MBA	**ChatGPT would pass an MBA degree exam at Wharton (UPenn).**
Accounting	**GPT-3.5 would pass the US CPA exam.**
Legal	**GPT-3.5 would pass the bar in the US.**
Medical	**ChatGPT would pass the United States Medical Licensing Exam (USMLE).**
AWS certificate	**ChatGPT would pass the AWS Certified Cloud Practitioner exam.**
IQ (verbal only)	**ChatGPT scores IQ=147, 99.9th %ile.**
SAT exam	**ChatGPT scores 1020/1600 on SAT exam.**

https://lifearchitect.ai/chatgpt/

GPT Evolution

Not only Bigger and Bigger



Fine Tuning

Or

In context Few shot Learning

175B parameters

96 decoders

GPT-3

2048 token size

Fine Tuning

1.5B

48 decoders

GPT-2

1024 token size

Fine Tuning

117M

12 decoders

GPT-1

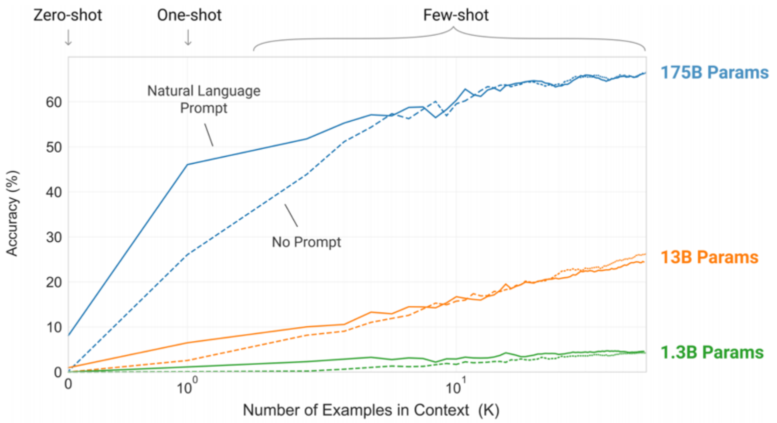
512 token size

2018 → 2020 → 2020

As the model and dataset get larger, it will know more and more

"GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model."

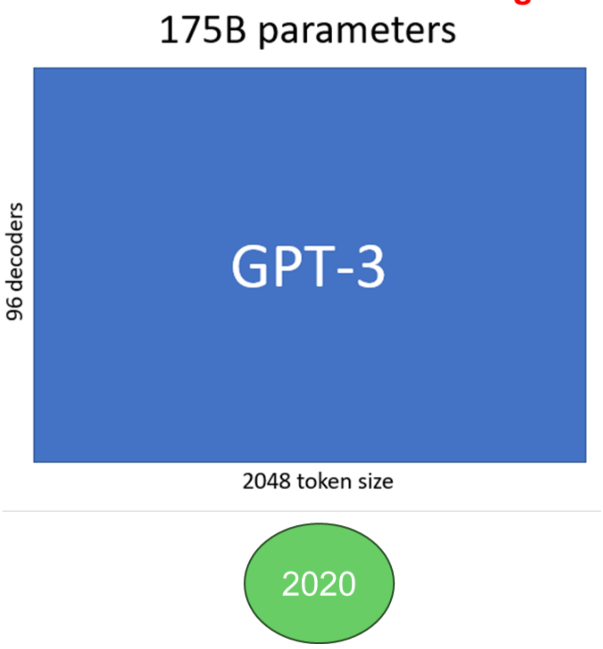
From **Language Models are Few-Shot Learners (2020)**



GPT Evolution

Not only Bigger and Bigger

Fine Tuning
Or
In context Few shot Learning

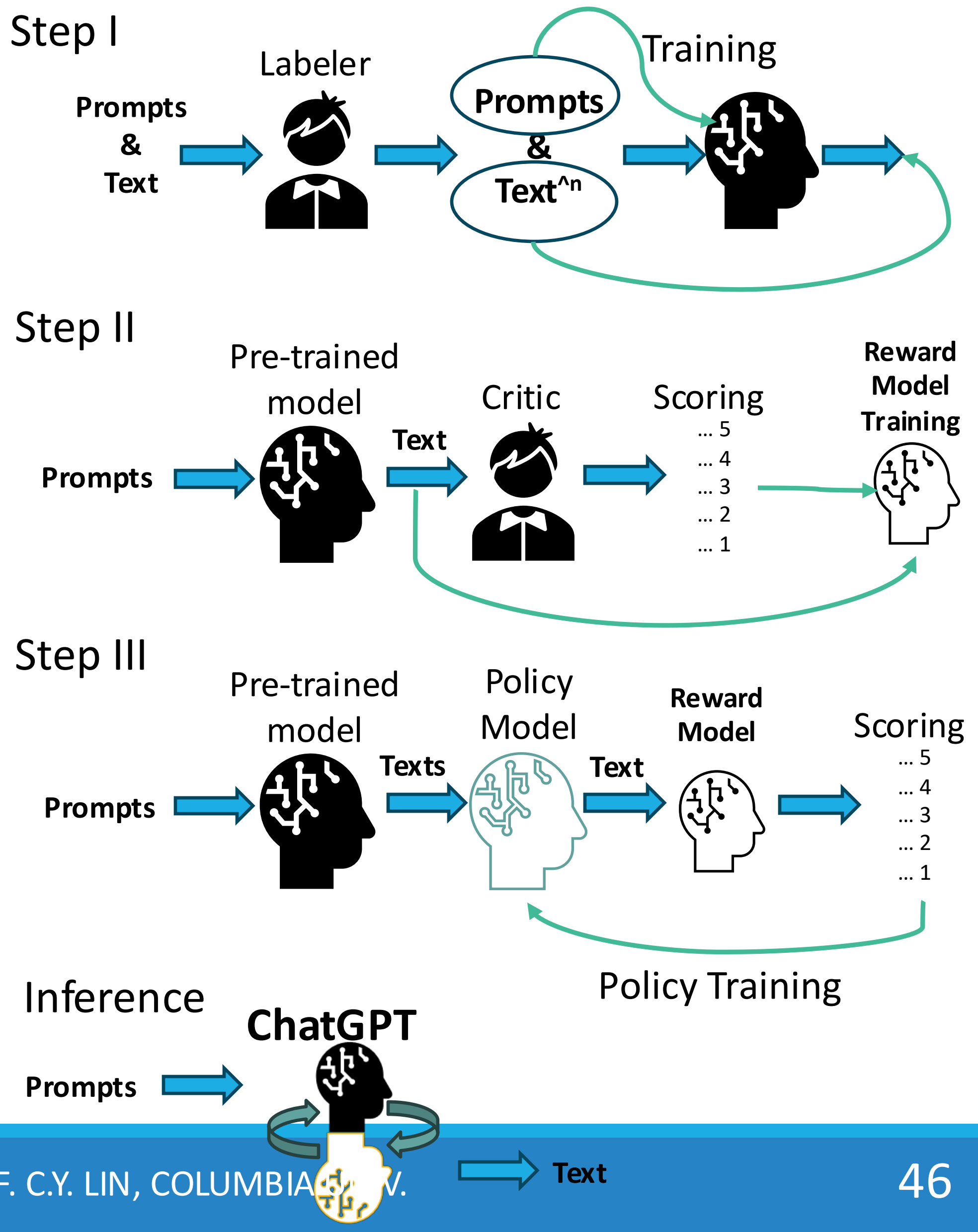


?

How does the Model Answer smartly or more like an Adult human

Thinking and Answering policy optimization

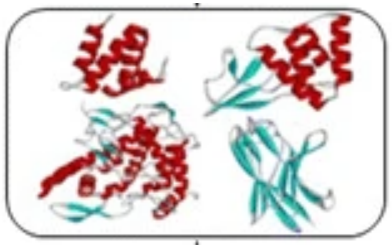
Reinforcement Learning from Human Feedback (RLHF)



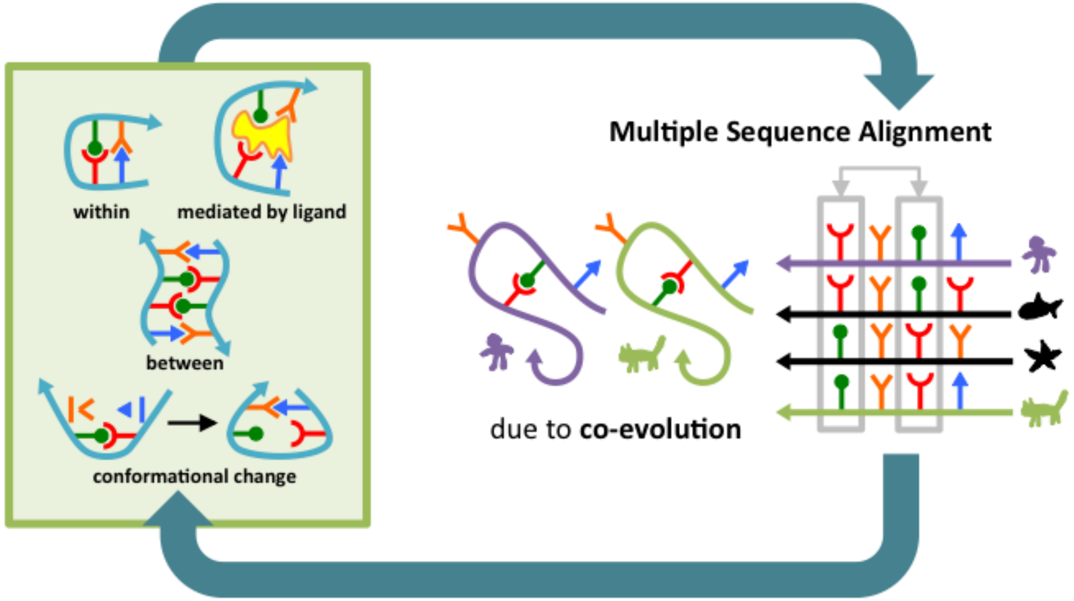
New Era is Now – AI is making Newton Moment of Life Science

The logics of Protein Structure Models

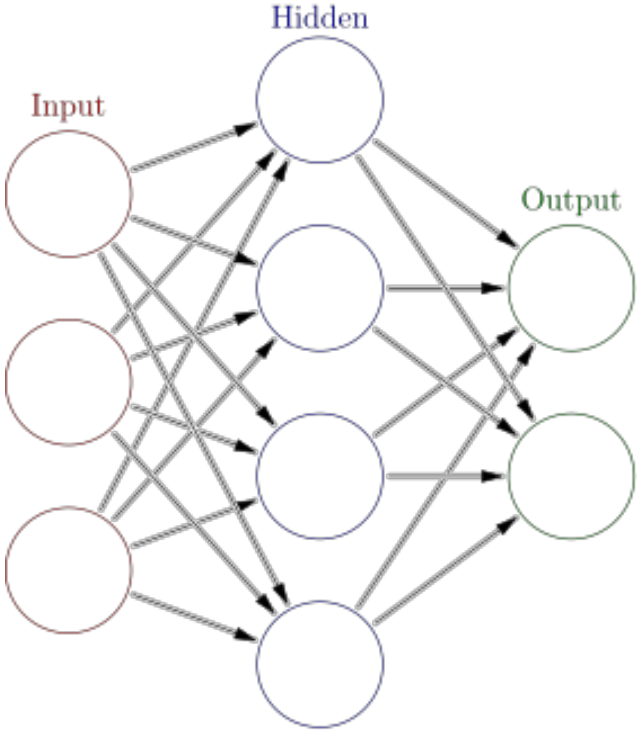
Template for similar secondary structure features



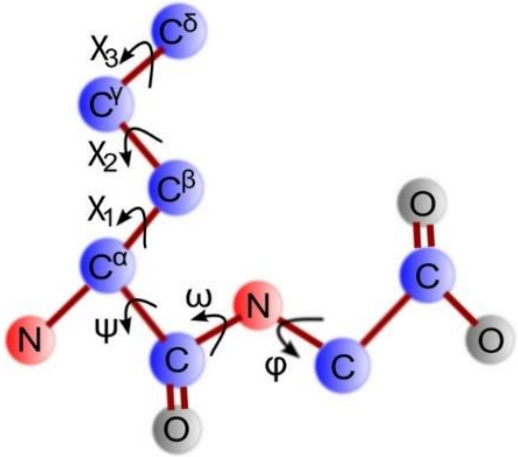
MSA for co-evolution features in different species



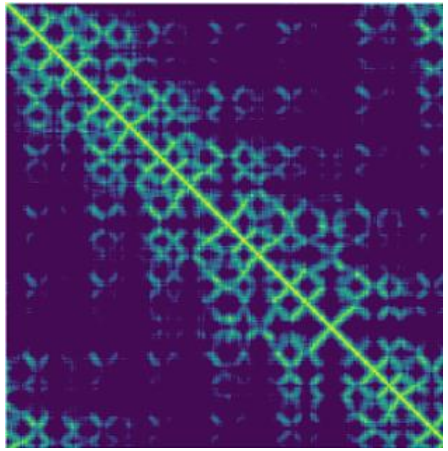
Model training



Orientation

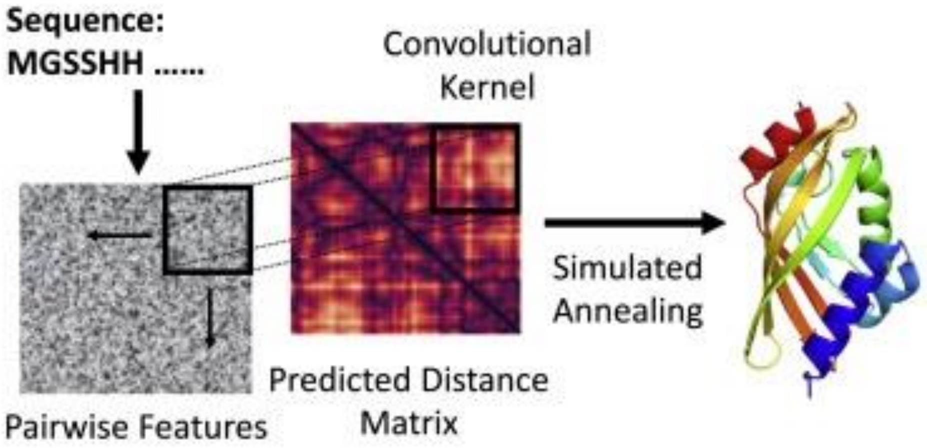


Distance Matrix

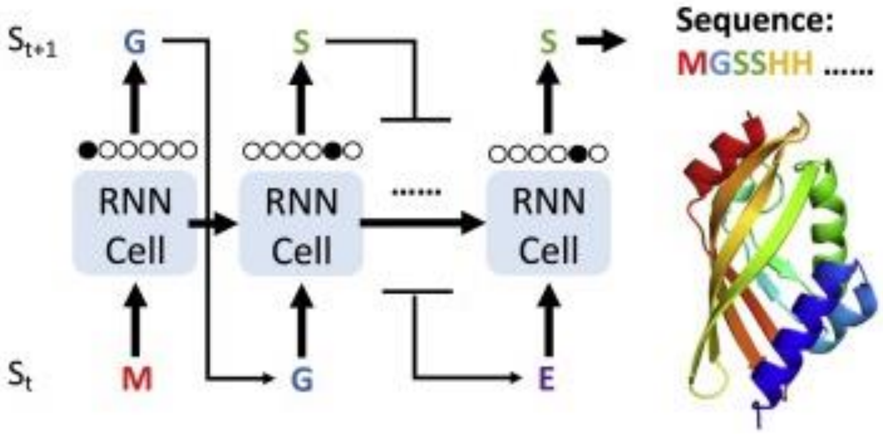


Deep Learning for Protein Structure Modeling

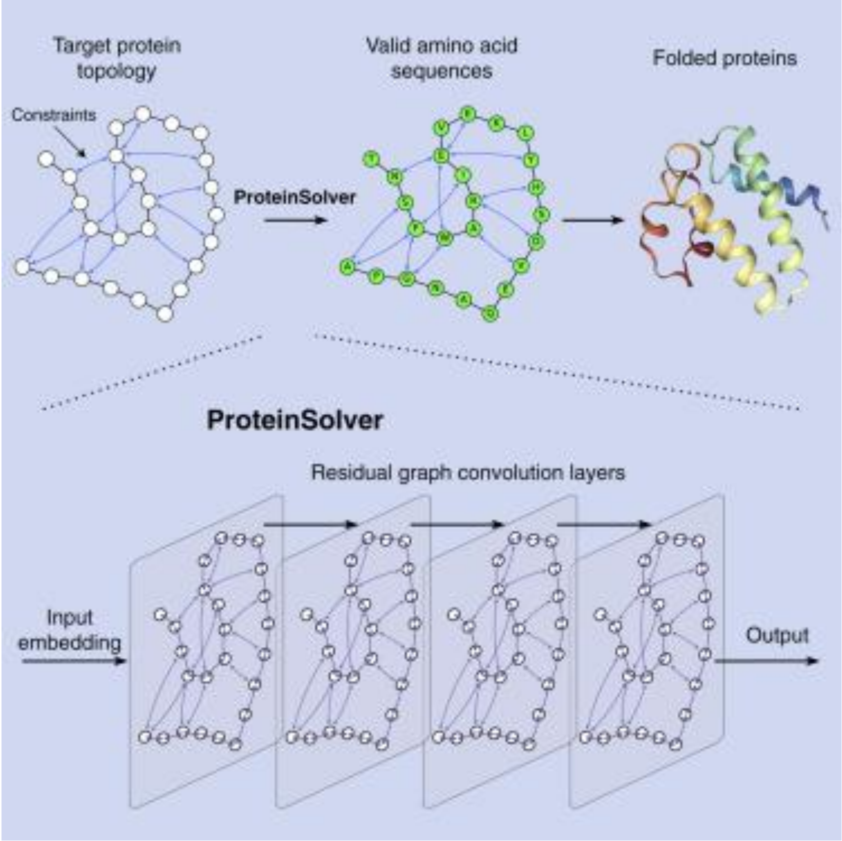
Convolution Neural Network



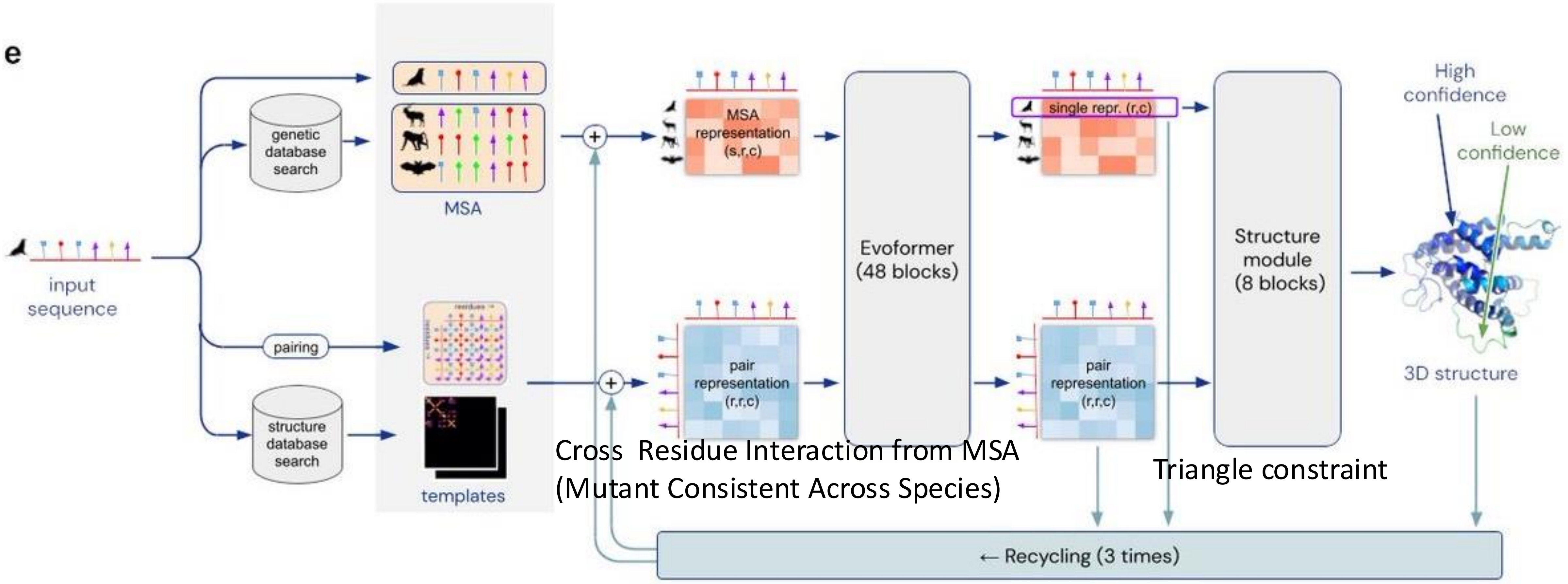
Recurrent Neural Network



Graph Neural Network



Alphafold2



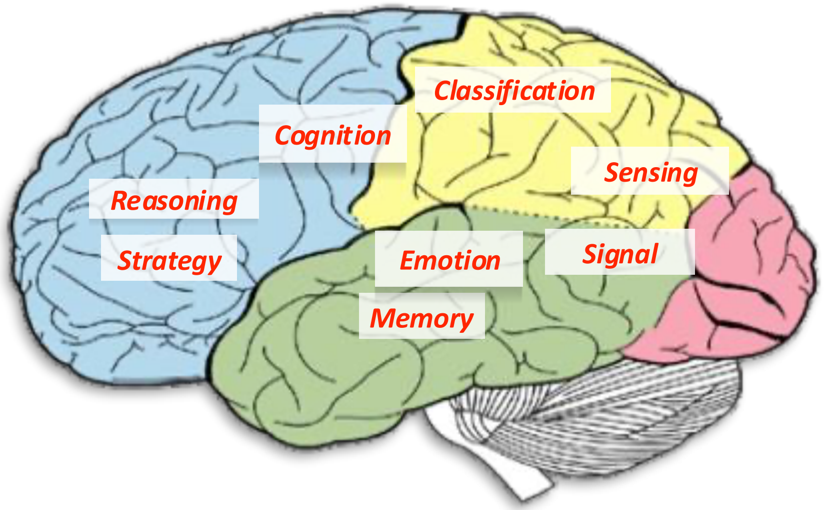
Cross Residue Interaction from MSA (Mutant Consistent Across Species)

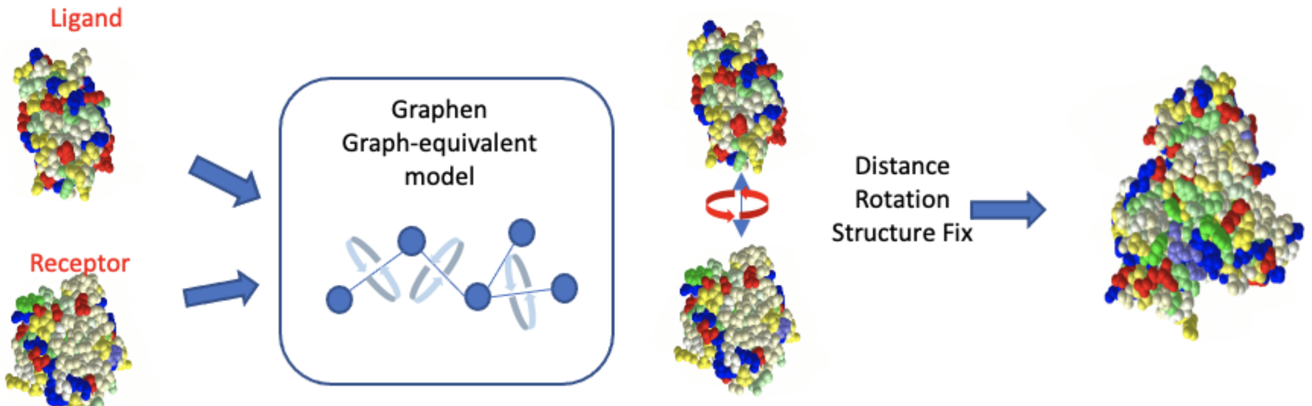
Triangle constraint

1. Training resources : 128 slides TPU v3 ≈ 300 slides GPU V100
2. Attention with Graph-Based Invariant Model Concatenate
3. Amber force Refine Side chains

Ref. https://deepmind.com/blog/article/AlphaFold-Using-AI-for-scientific-discovery

Graphen Atom Tools that better simulate biological functions



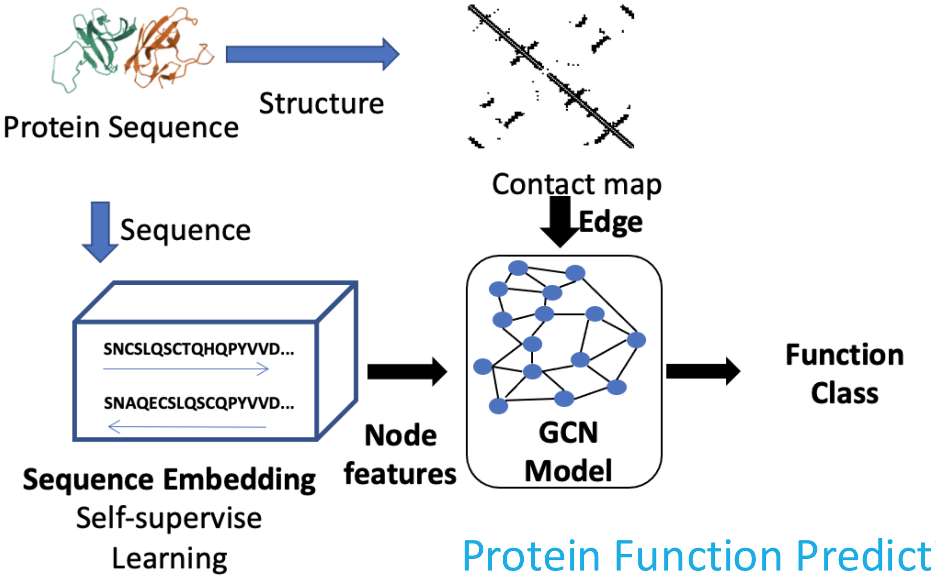


Atom Network Modeling

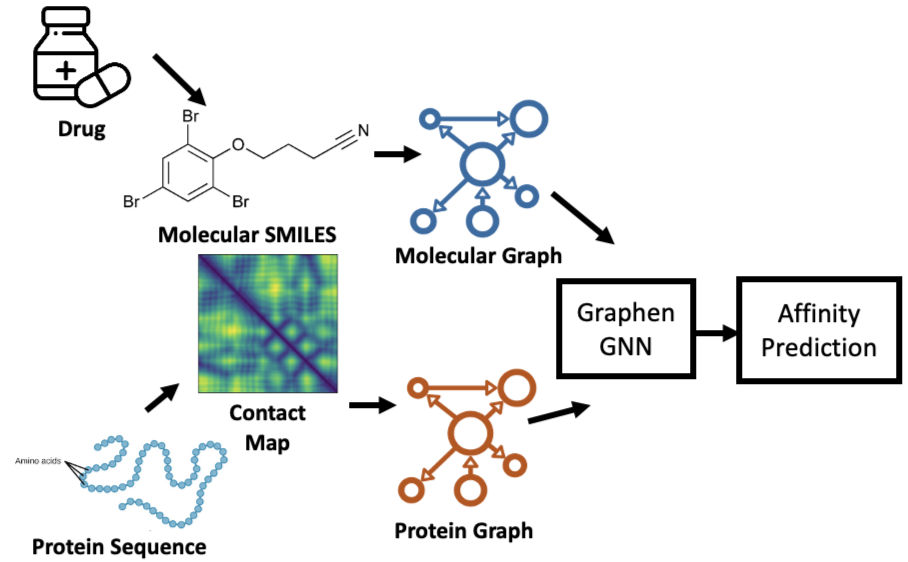
https://www.graphen.ai/products/ardi.htm

Ardi Graph Analytics and Database of zillions of nodes and edges were deployed in one of the world's largest institutes in 2018.

Graphen's Differentiator:
Graph Computing + Deep Learning + Generative AI ➔ End-to-End New Drug Design



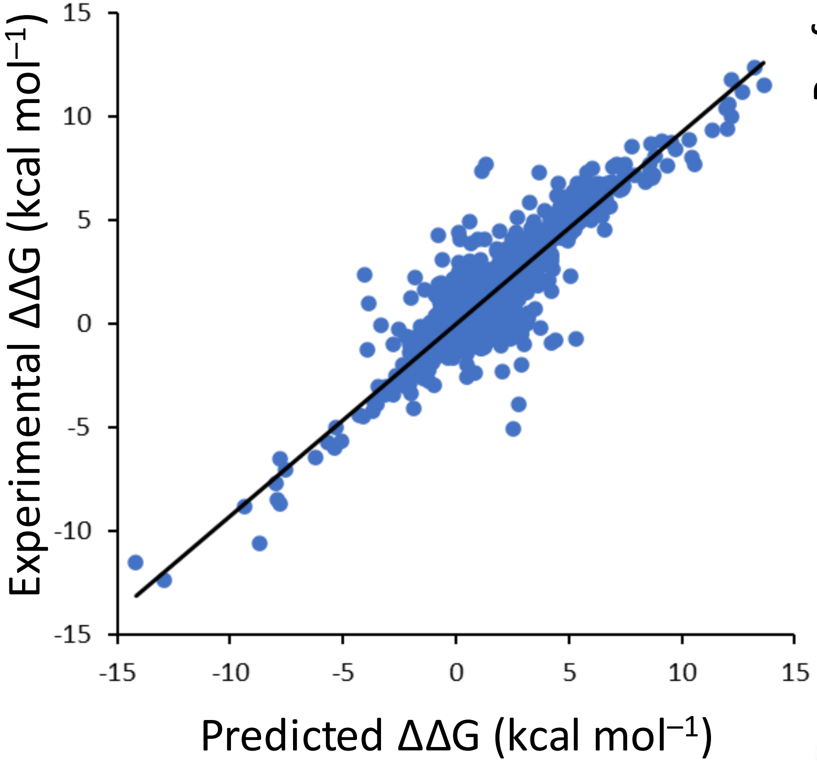
Protein Function Prediction

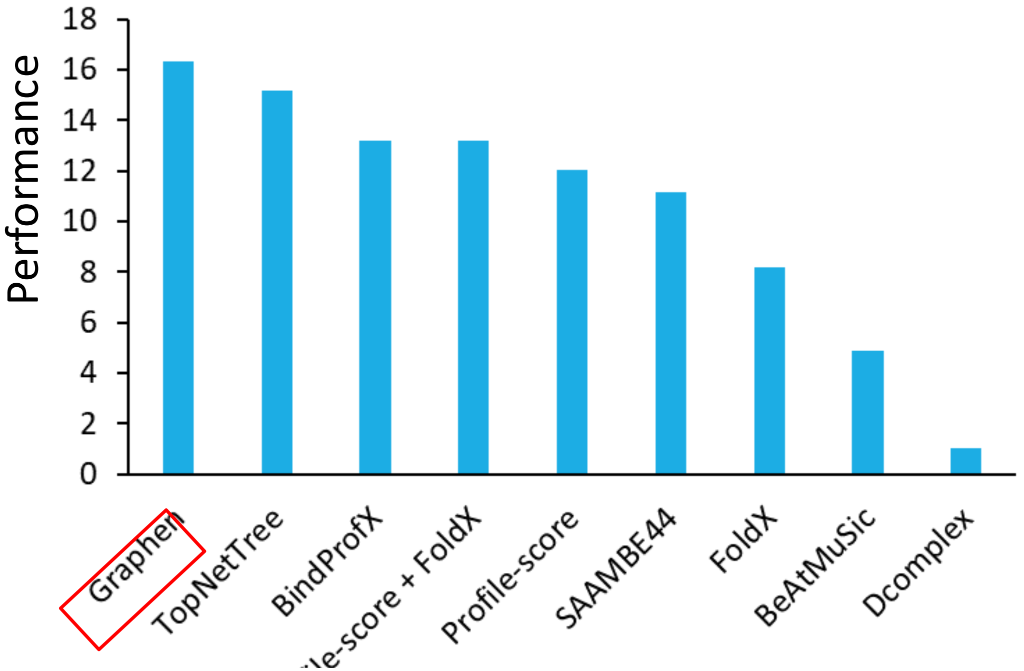


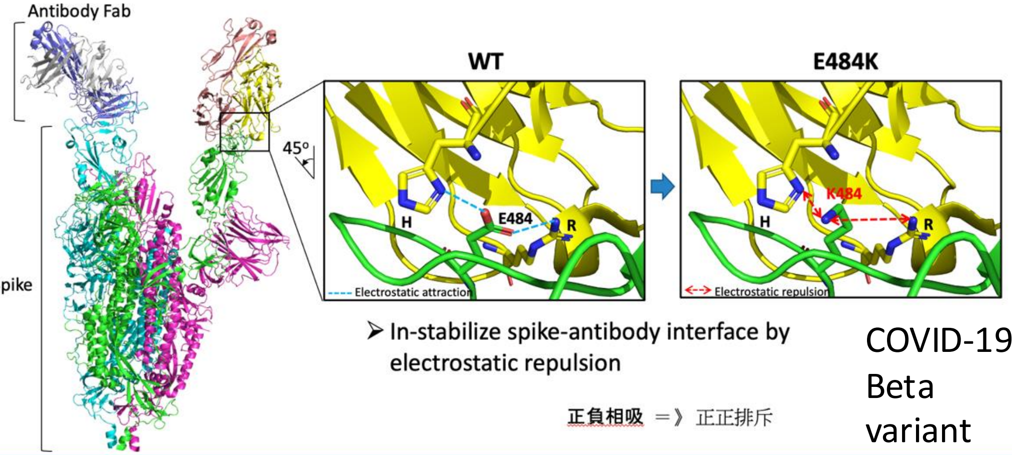
Affinity Prediction

Example: Graphen binding energy prediction model is better than the current state-of-the-art in predicting ΔΔG

Models	*Pearson's r*
Graphen-Atom	0.914
TopNetTree	0.850
BindProfX	0.738
Profile-score + FoldX	0.738
Profile-score	0.675
SAAMBE44	0.624
FoldX	0.457
BeAtMuSic	0.272
Dcomplex	0.056







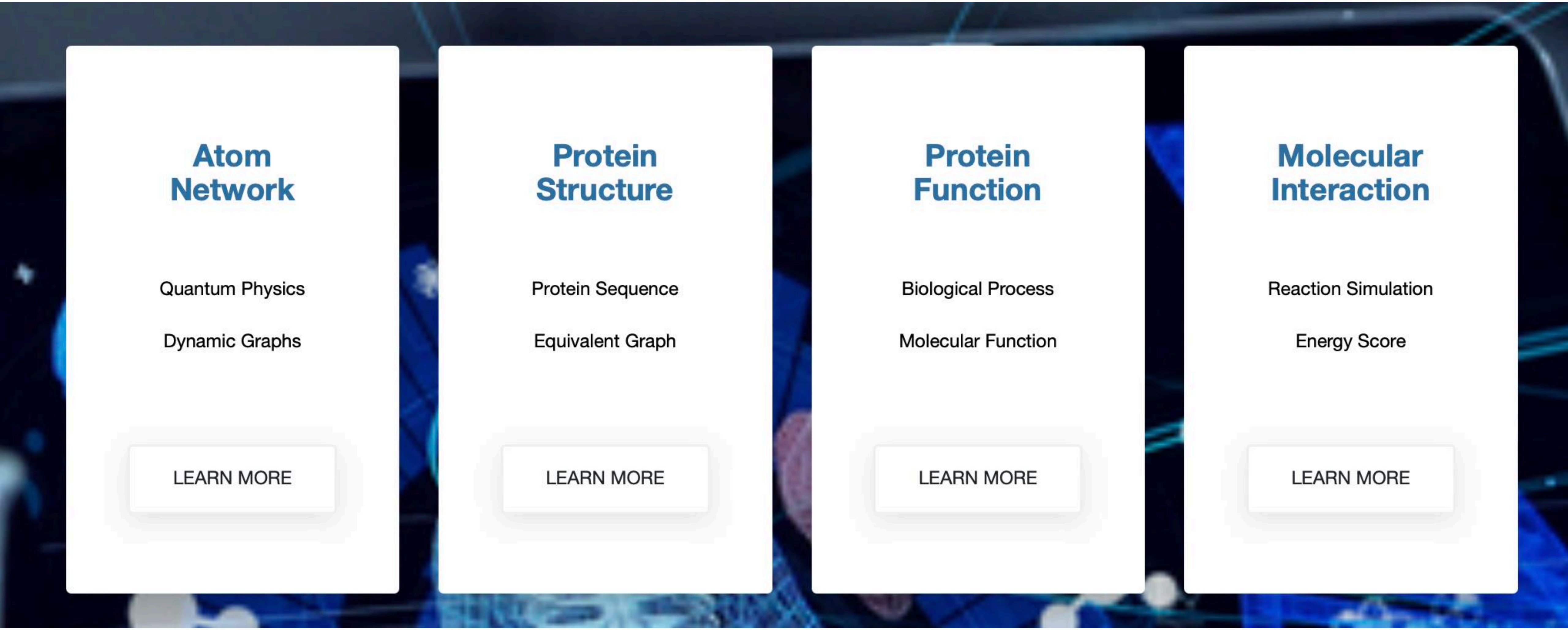
- In-stabilize spike-antibody interface by electrostatic repulsion

正負相吸 =》 正正排斥

COVID-19 Beta variant

Graphen Atom Toolkits - I

https://www.graphen.ai/products/atom.html



Atom Network

Quantum Physics

Dynamic Graphs

LEARN MORE

Protein Structure

Protein Sequence

Equivalent Graph

LEARN MORE

Protein Function

Biological Process

Molecular Function

LEARN MORE

Molecular Interaction

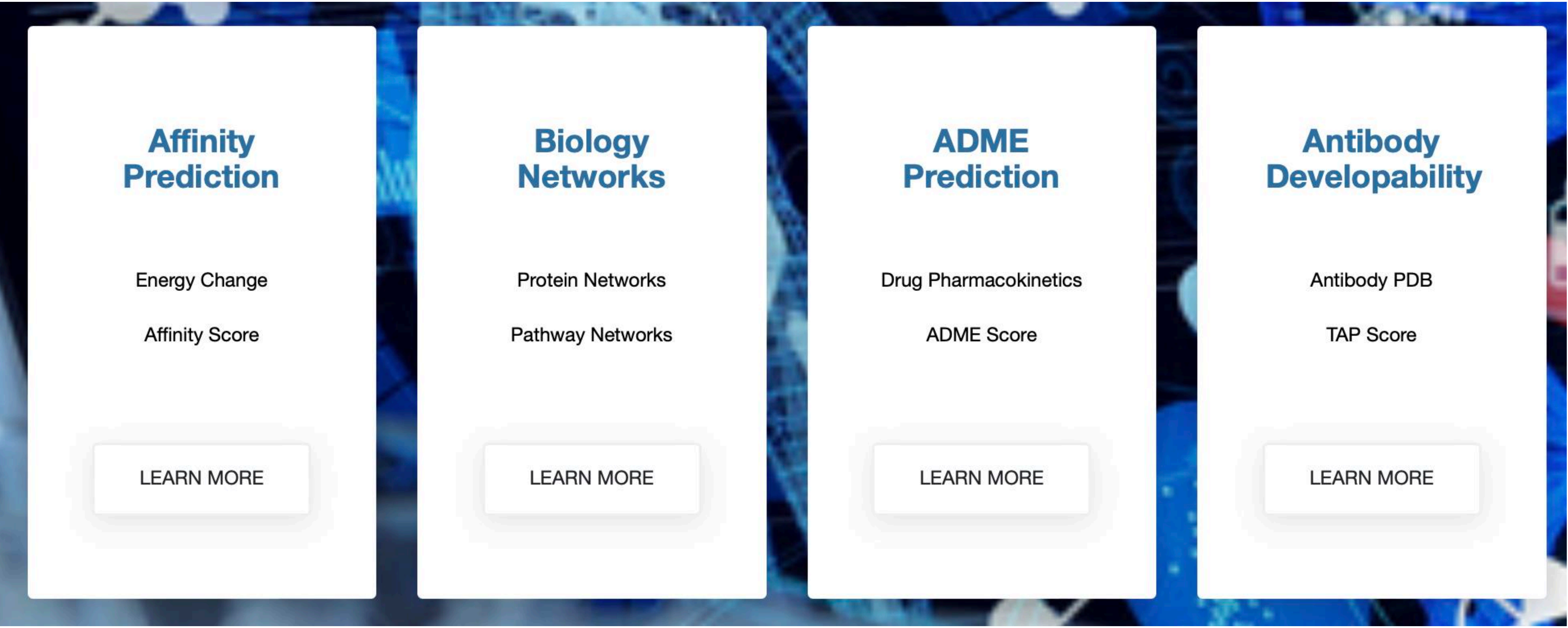
Reaction Simulation

Energy Score

LEARN MORE

Graphen Atom Toolkits - II

https://www.graphen.ai/products/atom.html



Affinity Prediction

Energy Change

Affinity Score

Biology Networks

Protein Networks

Pathway Networks

ADME Prediction

Drug Pharmacokinetics

ADME Score

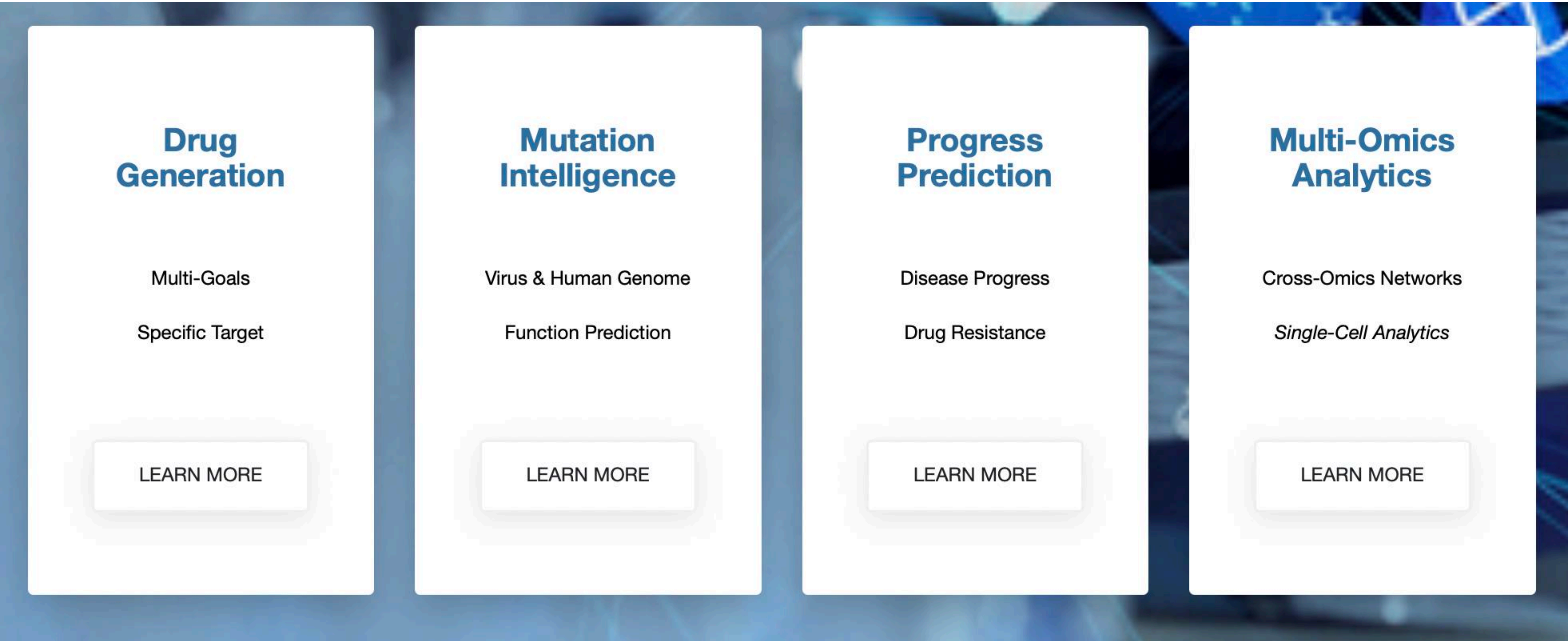
Antibody Developability

Antibody PDB

TAP Score

Graphen Atom Toolkits - III

https://www.graphen.ai/products/atom.html



Drug Generation

Multi-Goals

Specific Target

Mutation Intelligence

Virus & Human Genome

Function Prediction

Progress Prediction

Disease Progress

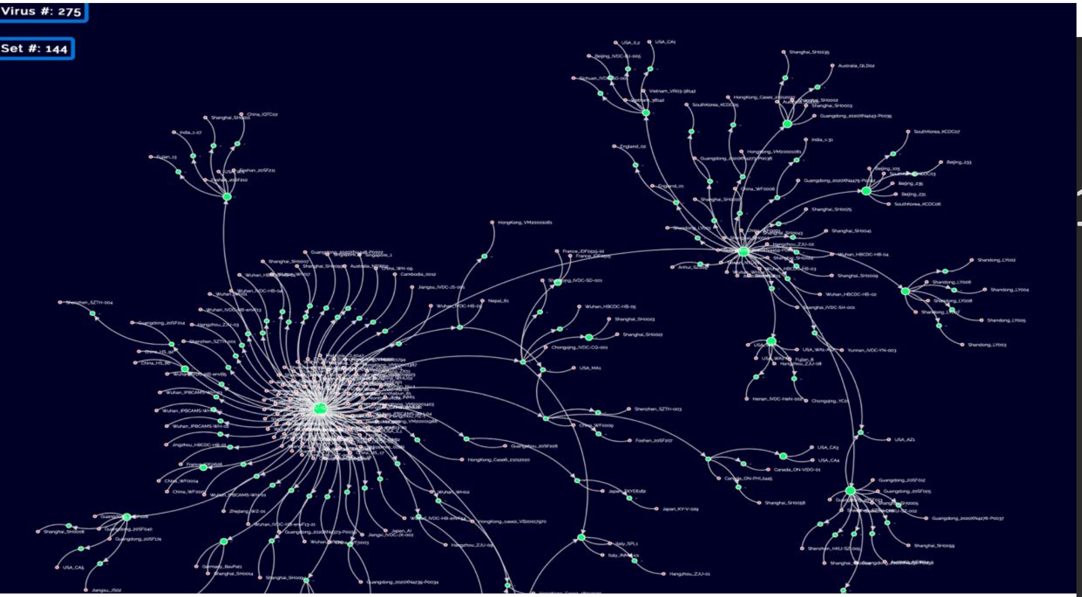
Drug Resistance

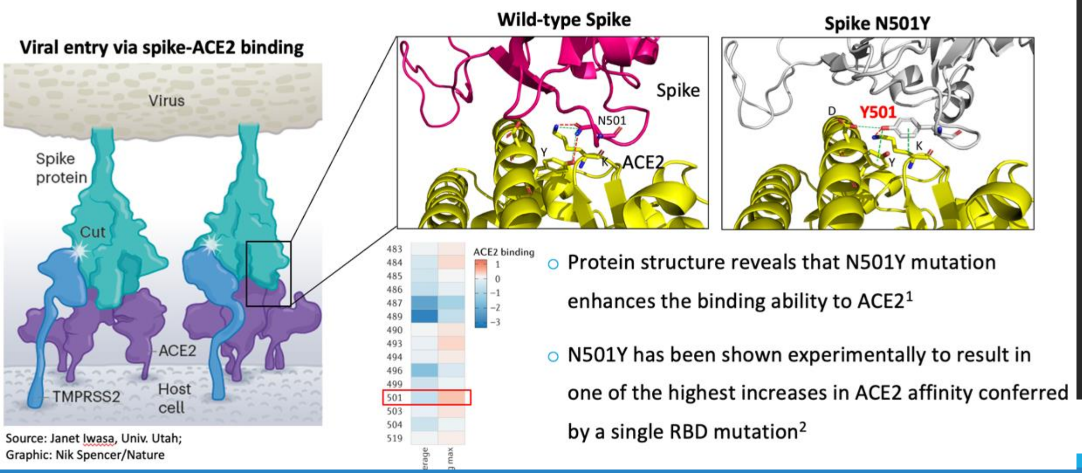
Multi-Omics Analytics

Cross-Omics Networks

Single-Cell Analytics

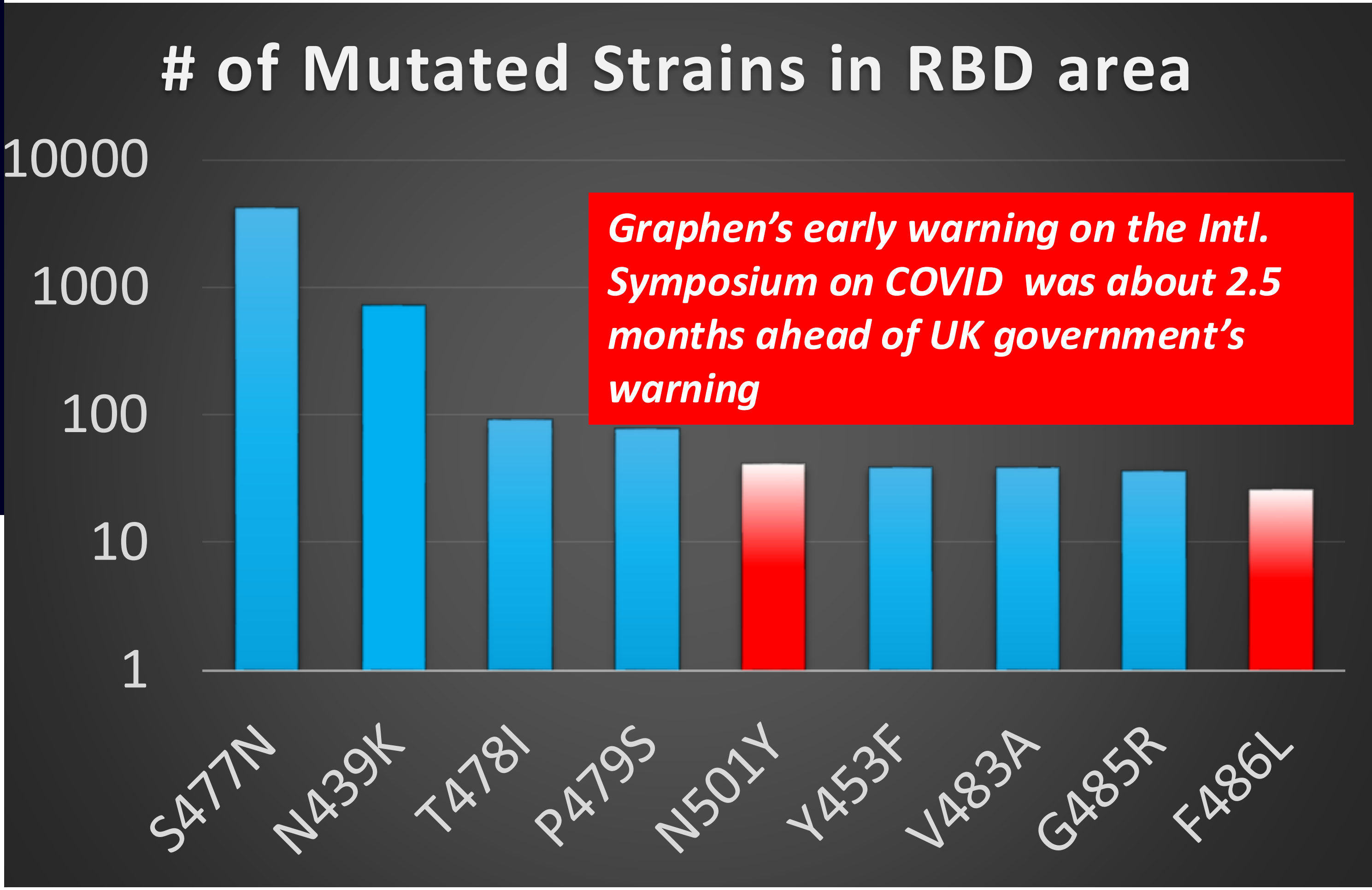
Graphen AI provided early warning of Alpha variants in September 2020, about 2.5 months ahead of first warning issued by the UK government





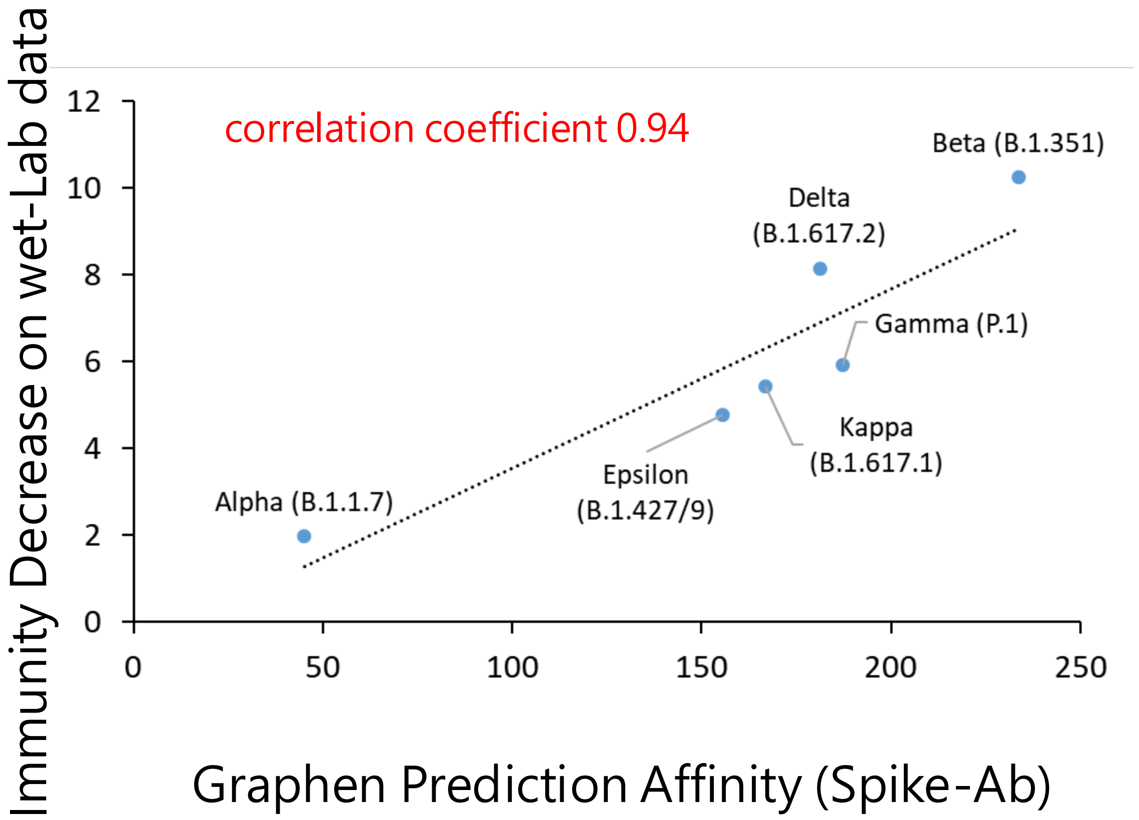
- Protein structure reveals that N501Y mutation enhances the binding ability to ACE2[1]
- N501Y has been shown experimentally to result in one of the highest increases in ACE2 affinity conferred by a single RBD mutation[2]

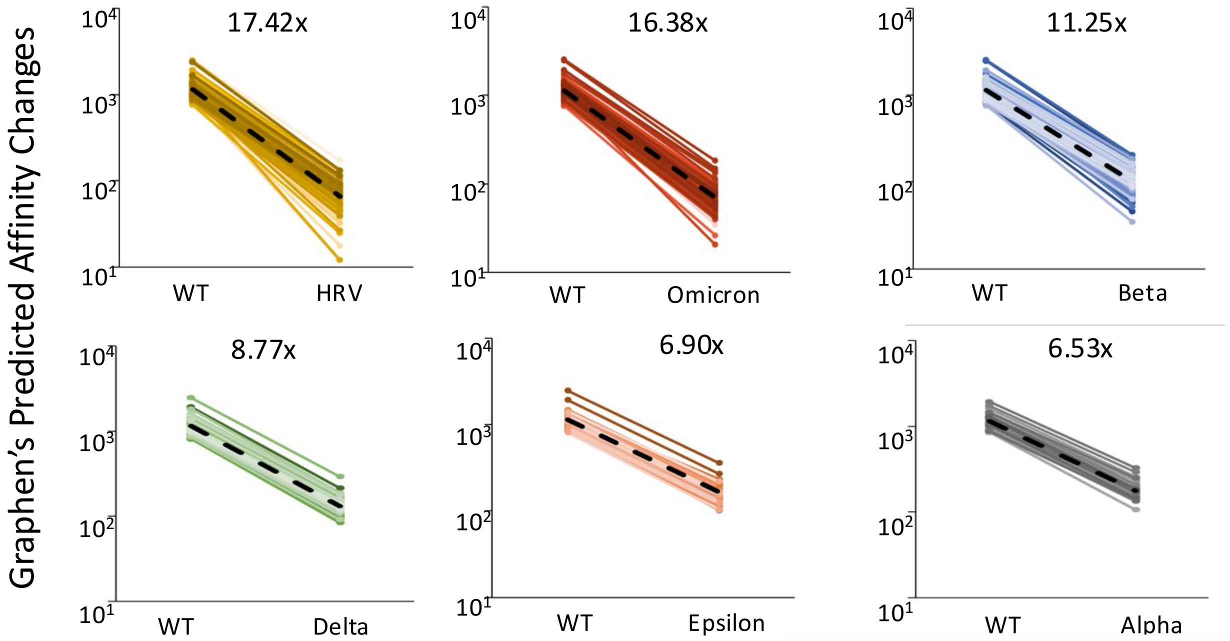
Source: Janet Iwasa, Univ. Utah; Graphic: Nik Spencer/Nature



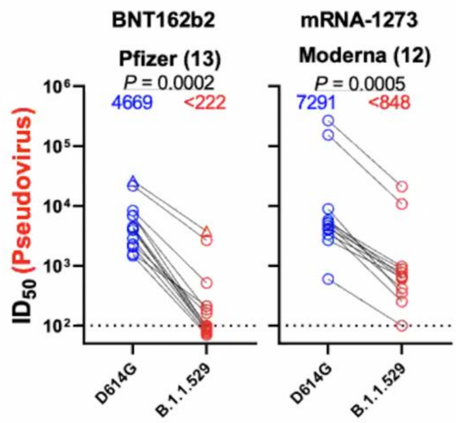
Graphen's early warning on the Intl. Symposium on COVID was about 2.5 months ahead of UK government's warning

Graphen’s prediction of variants’ functions (perfectively) matched real-world wet-lab data

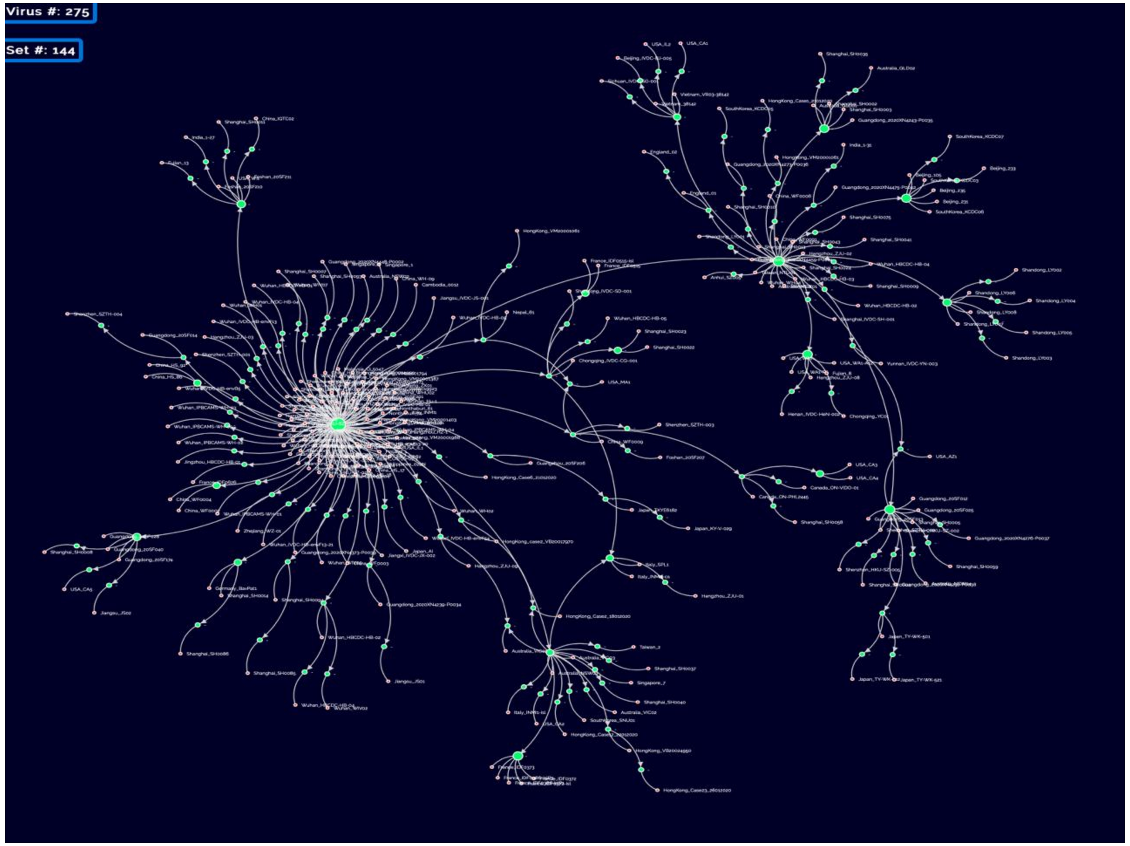




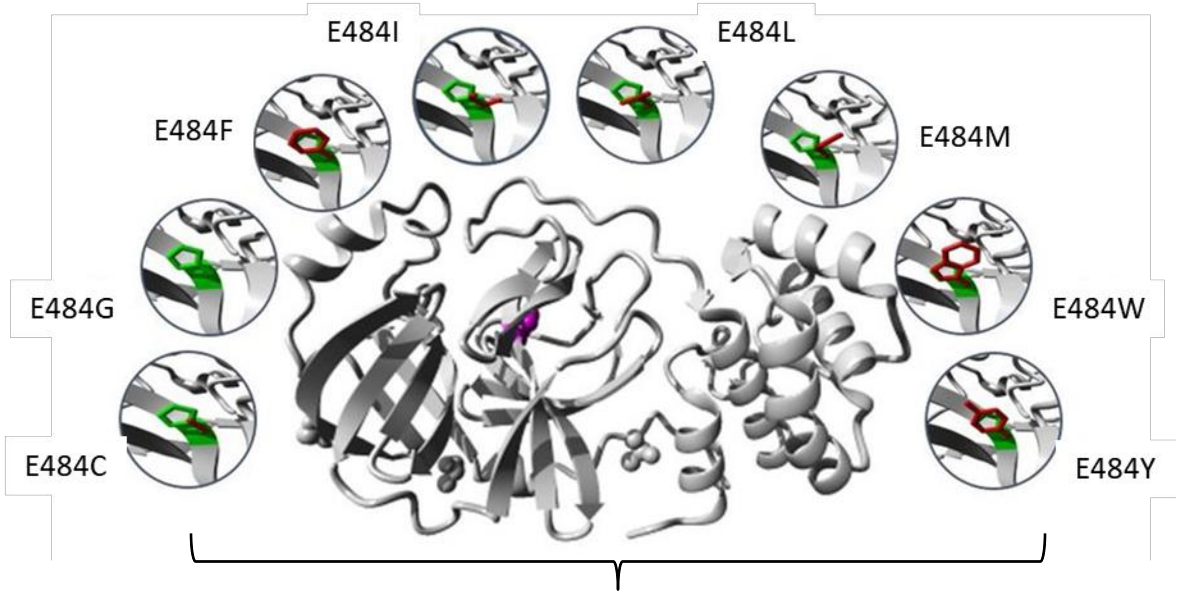
(Nov 26, 2021) Graphen predicted Omicron’s immunity escape is 16.38x. (Dec 15, 2021) Columbia Univ Med School presented the vaccine efficacy is down 20x by Pfizer and 9x by Moderna.



Graphen Designed “Future Vaccine” ： How can a virus mutate？What will happen after the mutation?？



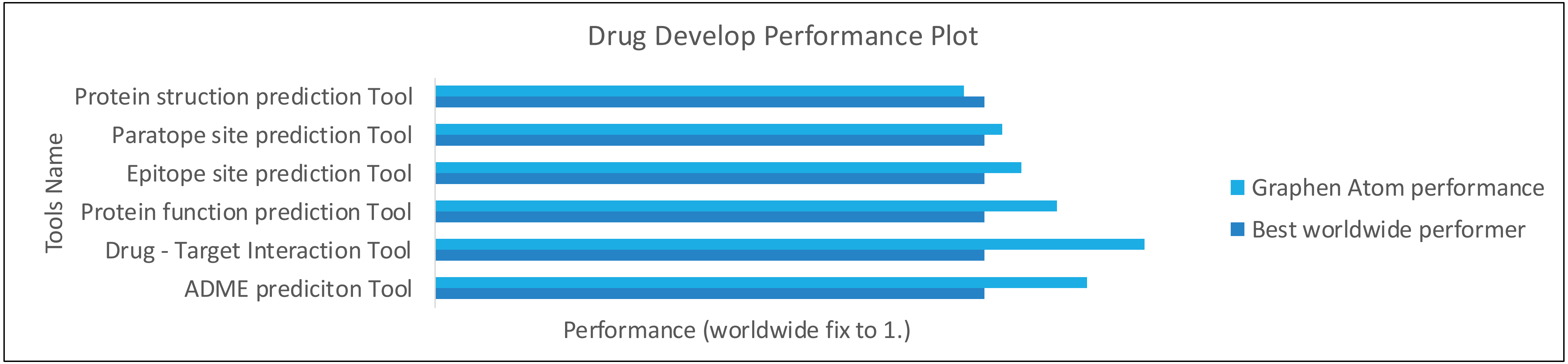
- 145 pair Spike Protein contact structure with antibody
- 296 pair Spike Residues of protein contact with antibodies
- **858,400 combinations in total**

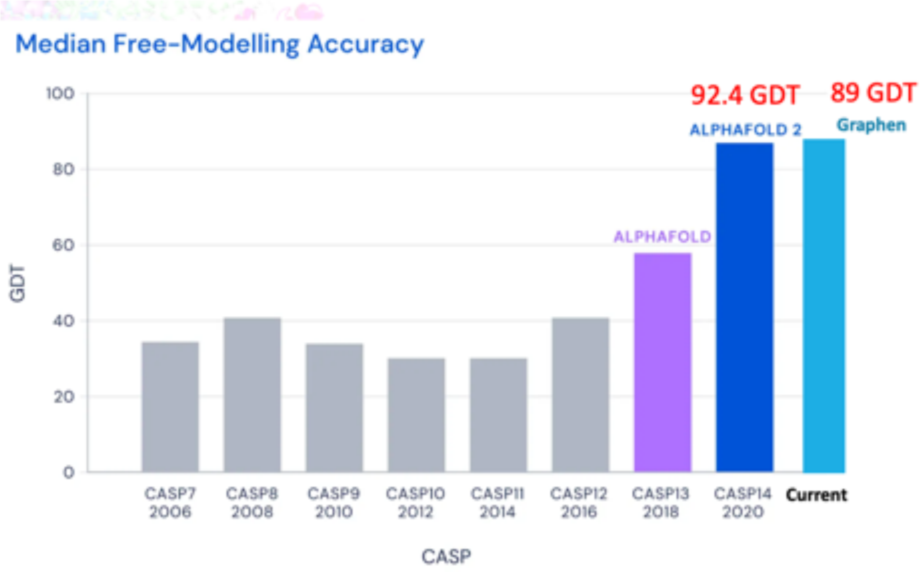


296 Contacting residues x 20 Amino acids x 145 structures

Graphen designed two “Future Vaccines” in September 2021 and October 2022 – Published in Nature Magazine’s *Scientific Reports* in Aug 2023 **scientific** reports **In silico prediction of immune-escaping hot spots for future COVID-19 vaccine design**

Graphen Atom: Best Performance on most key tools for AI Drug Development





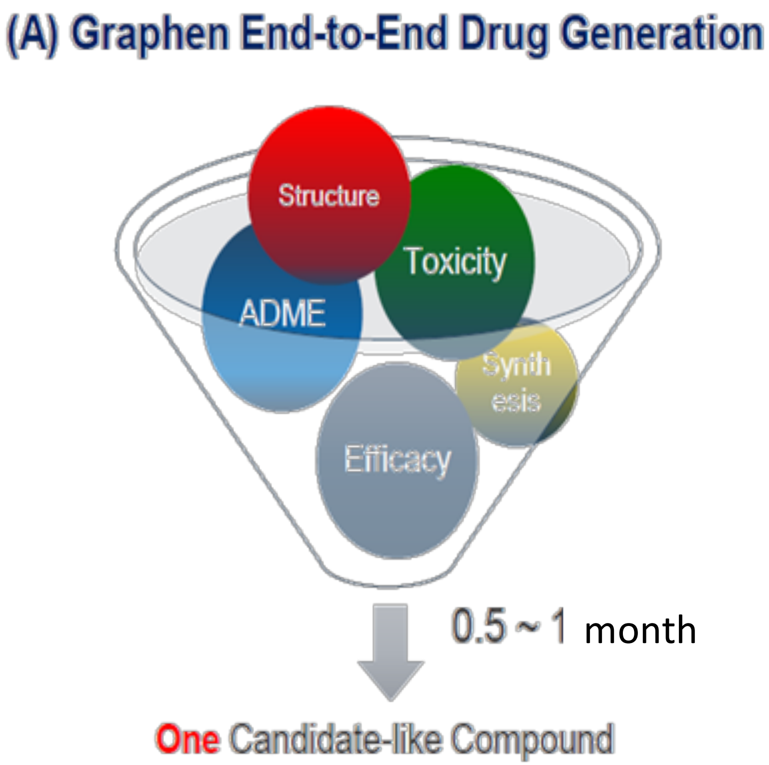
- ○ Graphen Atom uses **much fewer computational hardware** with similar performance to Alphafold2: ~1/50
- ○ Graphen Atom has **much more tools for drug development** than Alphafold2, which is limited to Protein Structure Prediction.

Tools is the most comprehensive (143 tools in 12 modules) and is the key to success

	Graphen	Isomorphic Labs (From Deepmind)	Potential Applications
Protein Structure Prediction	GDT Score : 89%	2024 Nobel Chemistry Prize GDT Score : 93%	Drug or Vaccine development
Molecular Co-Structures Simulation	Small molecular – Targeting Protein Antibody – Targeting Protein Nucleotide – Targeting Protein	None	Drug or Vaccine development
Drug Synergy & Side-Effect Prediction	Small molecular – Small molecular TherpeAntibody	None	Small molecular Drug development
ADMET Prediction	High Performance	None	Small molecular Drug development
Multi-objective Molecular Generator	Simulate High quality Small molecular drug (pLogP, permeability, QED score, affinity)	None	Small molecular Drug development
Molecular Retro-Synthesis	Based on USPTO reaction chain	None	Small molecular Drug development
Affinity Prediction	Antibody Affinity : 0.1 abnagtive log PKa	None	Antibody drug development
Solubility Prediction	Small molecular: pLogP Antibody drug: aggerate, SASA	None	Small molecular Drug development Antibody drug development
Permeability Prediction	Predict permeability in different cell line	None	Small molecular Drug development
DRL CDR Maturation	Based on affinity to maturate CDR domain	None	Antibody drug development
TAP Prediction	Based on affinity prediction to maturation CDR domain	None	Antibody drug development

Graphen's Unique AI Drug Generation System

Graphen

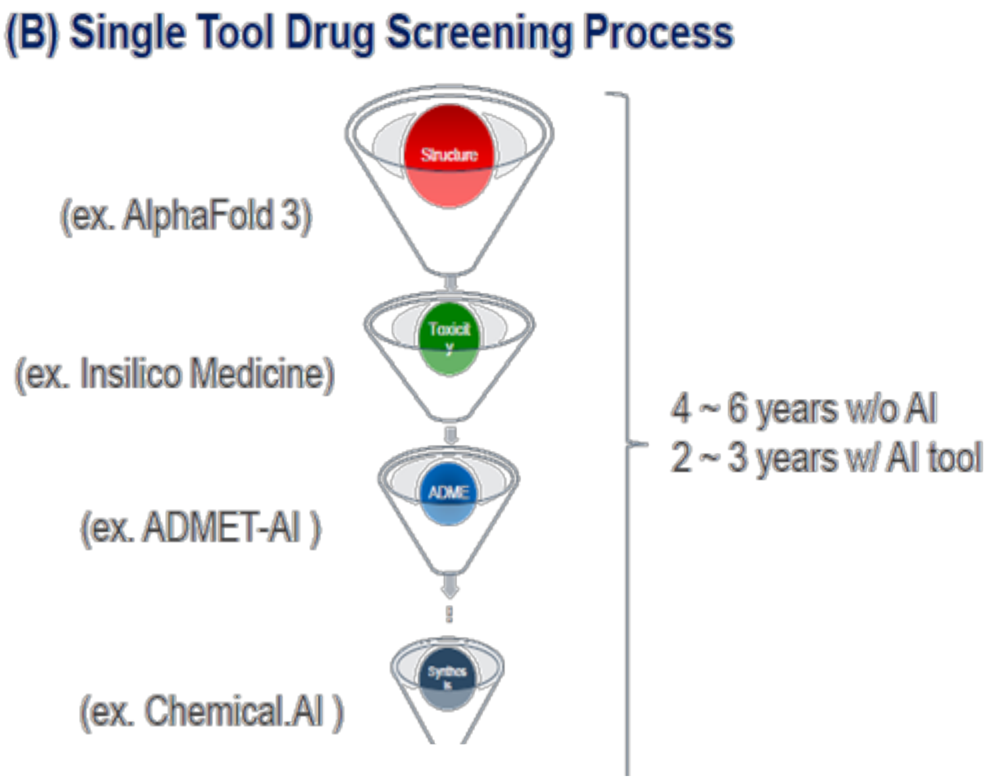

(A) Graphen End-to-End Drug Generation
Structure
Toxicity
ADME
Synthesis
Efficacy
0.5 ~ 1
One Candidate-like Compound

month

- AI "Virtual Experts" Work Together to create the best compound

To the best of our Knowledge: Success Rate: >90%

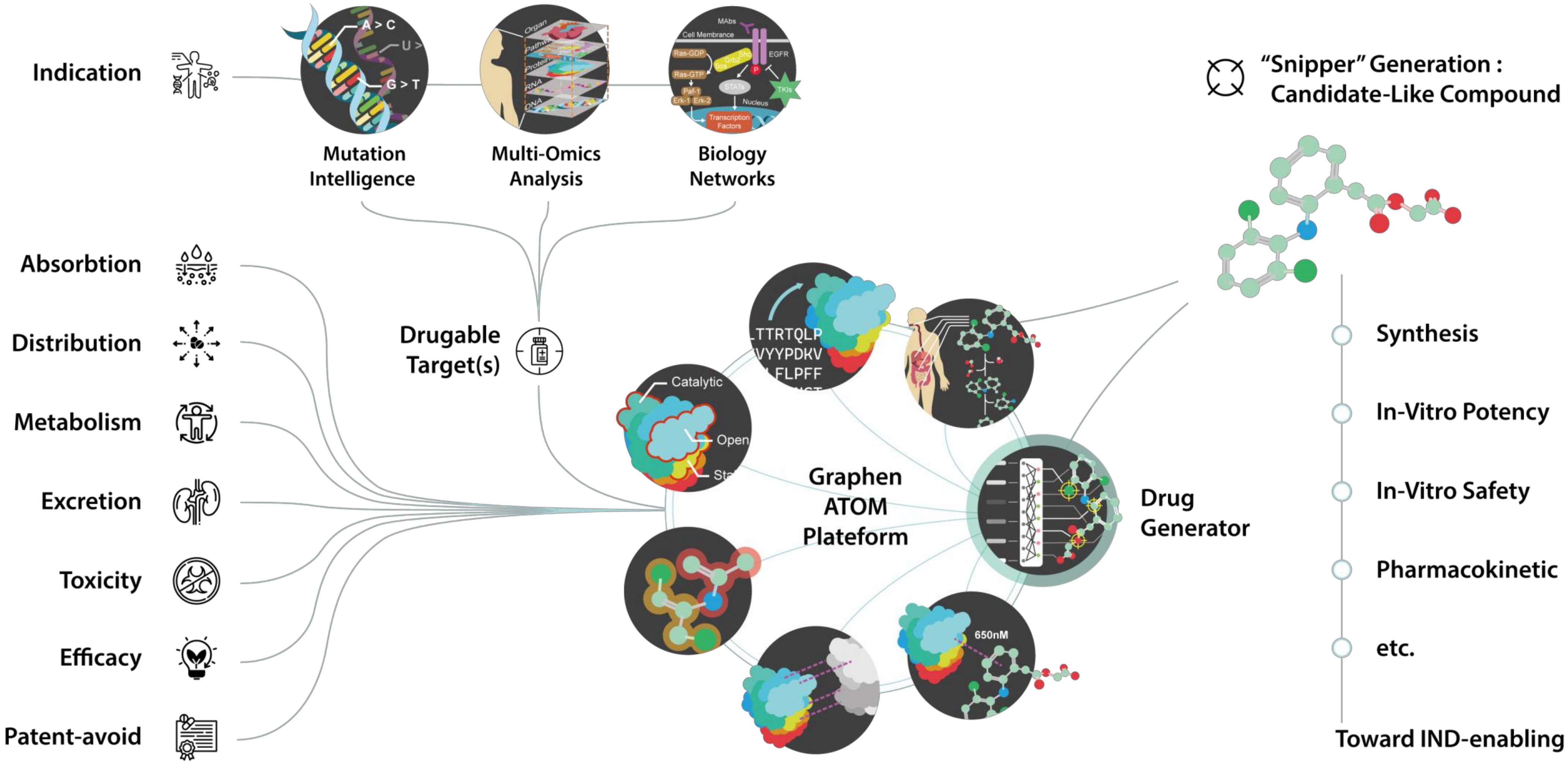
Others



- Human-AI Collaboration to make the process efficient

To the best of our Knowledge: Success Rate: <30%

Graphen Atom Platform : AI designing “ultimate” drugs



Indication

Mutation Intelligence

Multi-Omics Analysis

Biology Networks

Absorbtion

Distribution

Metabolism

Excretion

Toxicity

Efficacy

Patent-avoid

Drugable Target(s)

Graphen ATOM Plateform

Drug Generator

“Snipper” Generation : Candidate-Like Compound

Synthesis

In-Vitro Potency

In-Vitro Safety

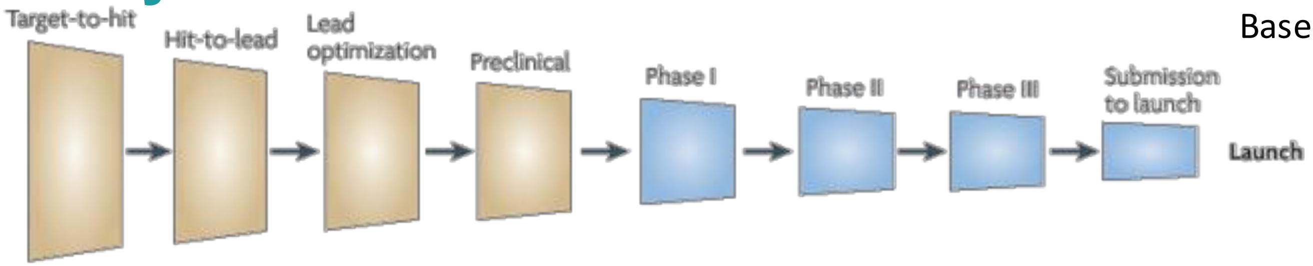
Pharmacokinetic

etc.

Toward IND-enabling

Comparisons of Cost & Time → 2 months vs ~15 years to create a verified effective drug compound; and is high effective, low side effects and low/no toxicity

Based on 2024 Dollar value



Cost & Time of General Pharma

	Target-to-hit	Hit-to-lead	Lead optimization	Preclinical	Phase I	Phase II	Phase III	Submission to launch	Launch
p(TS)	80%	75%	85%	69%	54%	34%	70%	91%	
WIP in Launch	6.65	5.32	3.99	3.39	2.34	1.26	0.43	0.3	0.27
Cycle Time (Year)	1.0	1.5	2.0	1.0	1.5	2.5	2.5	1.5	
Cost per launch (capitalized, M)	$37	$66	$164	$60	$108	$127	$122	$19	
	$267			$60	$376				

Total $703

Graphen Drugomics Cost & Time

	Target-to-hit	Hit-to-lead	Lead optimization	Preclinical	Phase I	Phase II	Phase III	Submission to launch	Launch
p(TS)	90.9%			80%	75%	60%	80%	95%	
WIP in Launch	1.1			1	0.8	0.6	0.36	0.3	0.27
Cycle Time (Year)	0.167			1.0	1.5	2.5	2.5	1.5	
Cost per launch (capitalized, M)	$0.2			$3	$37	$60	$102	$18	
	$0.2			$3	$217				

Total $220

Graphen Drugomics's **Total cost saving** of **$483M** per 0.27 launch *($703M vs $220M)*

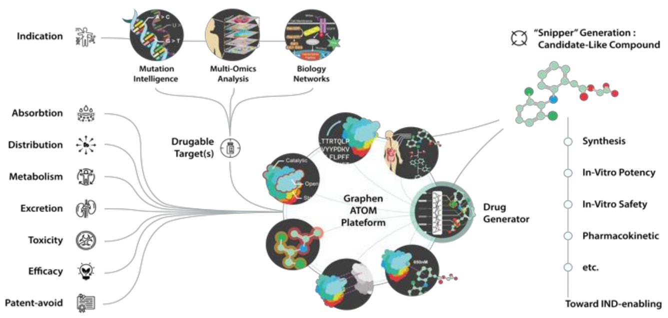
Graphen Drugomics's **IND-readies cost saving** of **$324M** per 0.27 launch *($327M vs $3.2M)*

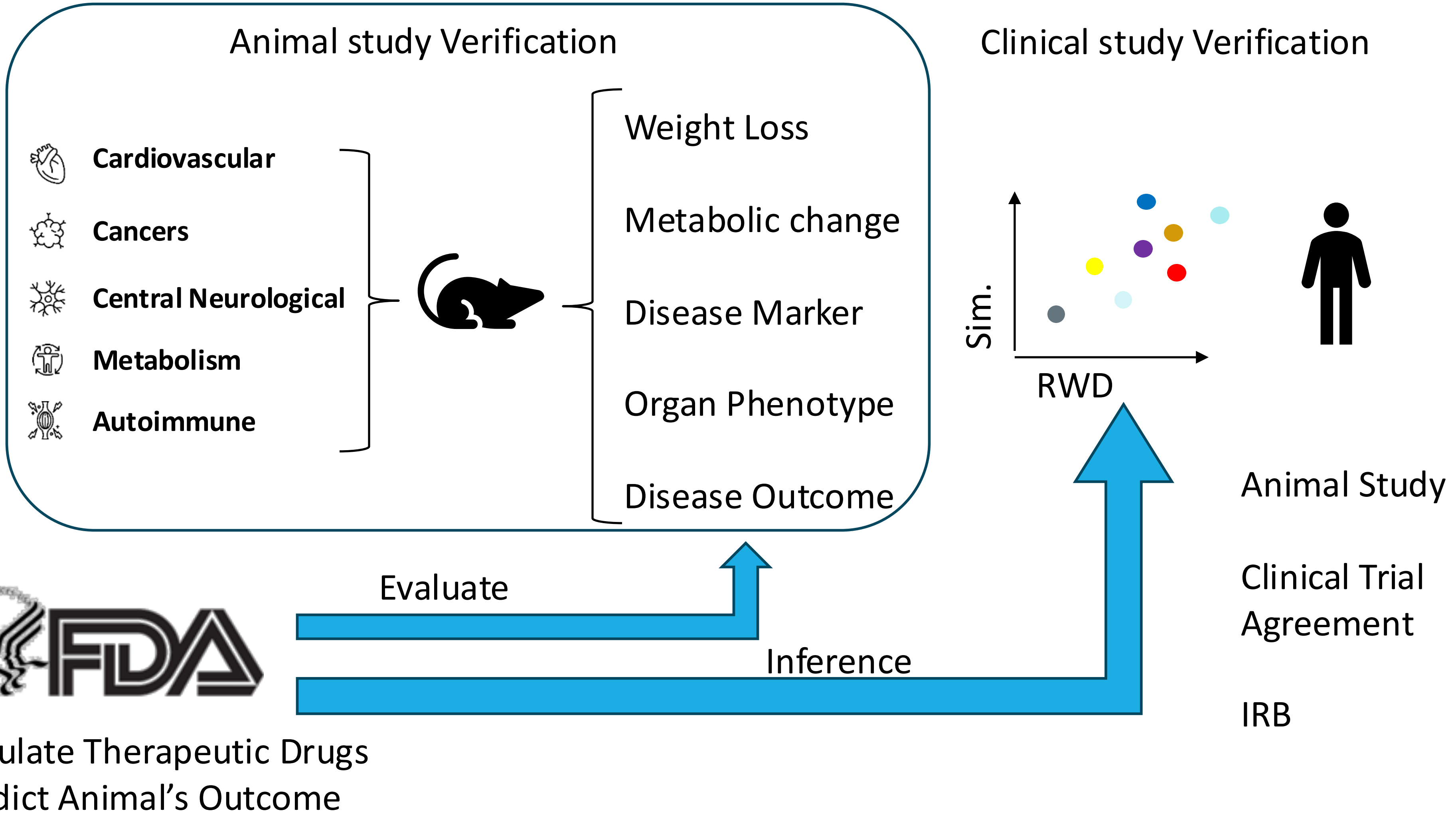
➔ Potential Leads: 1/1335 of Cost & 27x faster. Up to IND-Ready: 1/102 of Cost & Time: 4.7x faster

Roadmap: Replacing Animal & Clinical Trials in the future

➔ **In the future, new drugs will be designed in a short time. Make rare disease drugs and personalized drugs dream come true!!**

Drug Multi-model





Animal study Verification

Cardiovascular

Cancers

Central Neurological

Metabolism

Autoimmune

Weight Loss

Metabolic change

Disease Marker

Organ Phenotype

Disease Outcome

Clinical study Verification

Animal Study

Clinical Trial Agreement

IRB

Evaluate

Inference

Simulate Therapeutic Drugs
Predict Animal's Outcome

Goal:

1. Reduction
2. Refinement
3. Replacement

Roadmap: Make the Personal Precision Drug dream a reality

COLUMBIA UNIVERSITY

The main challenges in personal drug development include:

- ensuring precise drug targeting to minimize side effects,
- designing representative clinical trials to evaluate efficacy and safety, and
- developing innovative sustainable production and supply solutions.

Today, Graphen enables precise drug targeting, particularly in cellular and gene therapy, reducing side effects.

Today, Graphen Atom simulates drug-protein interactions and assess the impact of genetic mutations on cellular pathways. This enables personalized precision treatment and personalized drug development by:

- identifying promising compounds early, and
- developing personalized plans

➔ improved patient treatment outcome.

Future, Graphen designs and creates Personal Precision Medicine Drug .