GENERATIVE AI DEMYSTIFIED - JADE

ADAM GRZYWACZEWSKI, SENIOR DEEP LEARNING DATA SCIENTIST

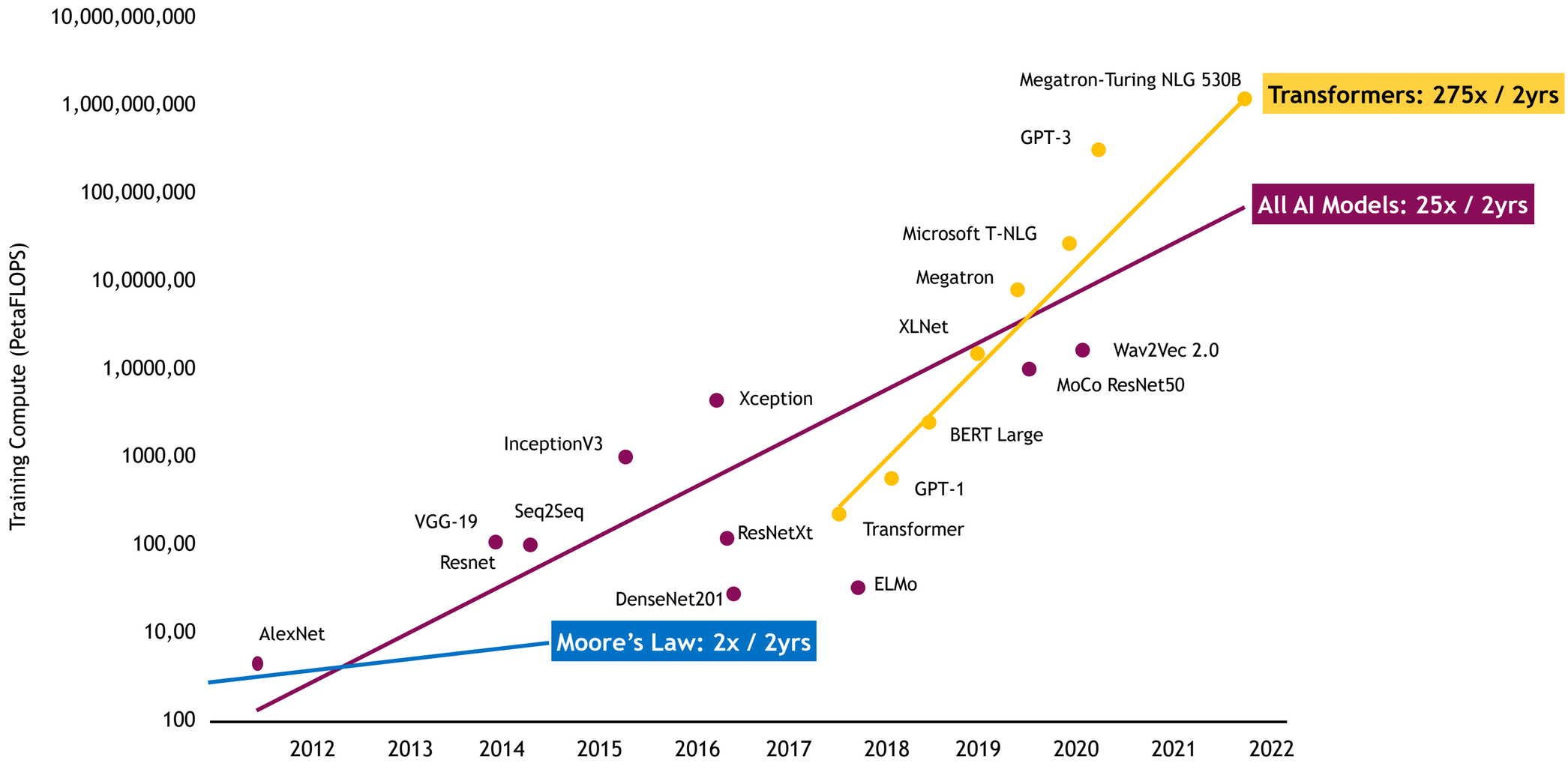
ABOUT ME

Adam Grzywaczewski – adamg@nvidia.com

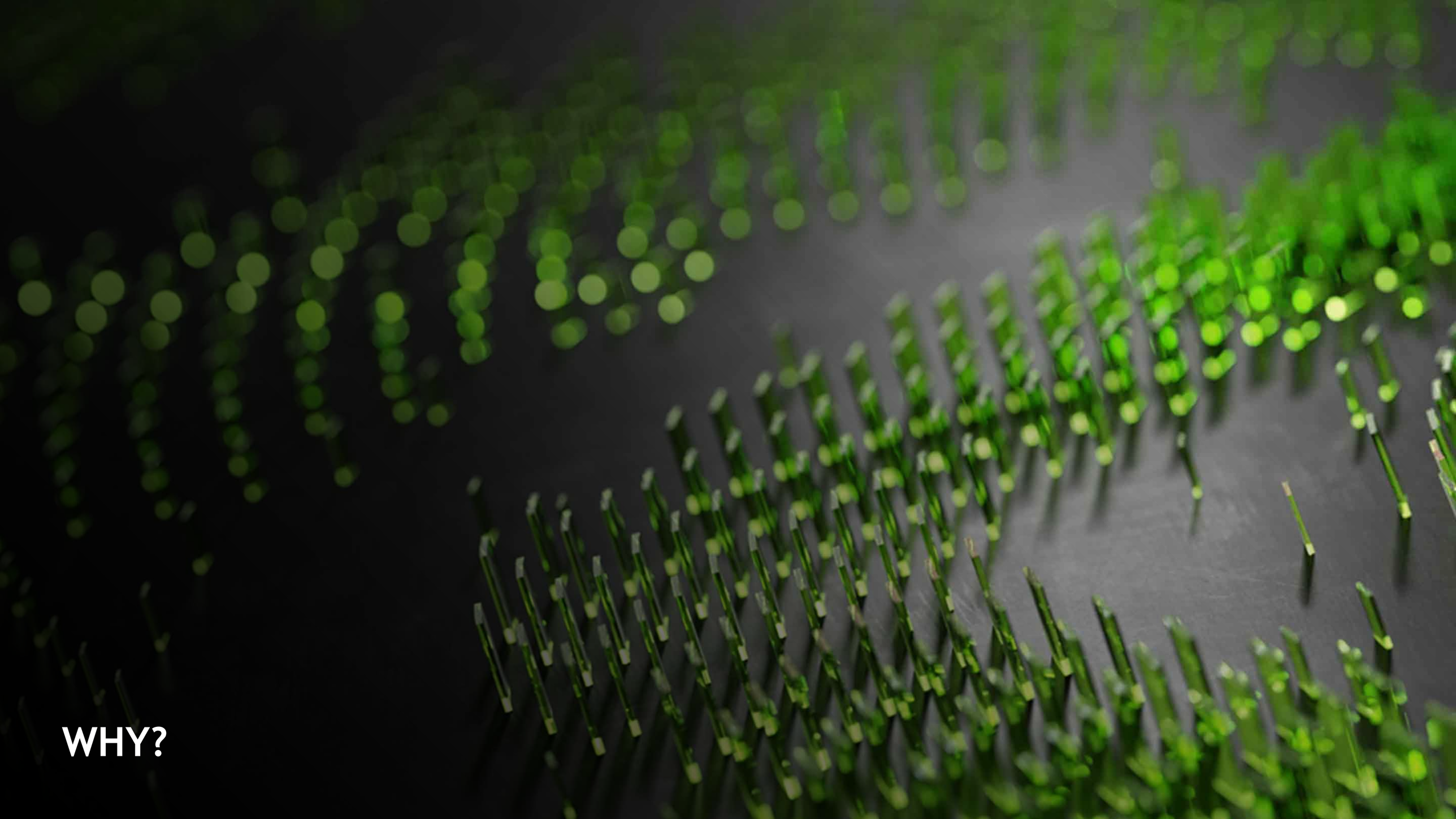
- Senior Deep Learning Data Scientist @ NVIDIA - Supporting delivery of AI / Deep Learning solutions
- Focusing on large scale/distributed training and efficient inference
- Expertise in Natural Language Processing

DRAMATIC INCREASE IN MODEL SIZES

The Trend Continues

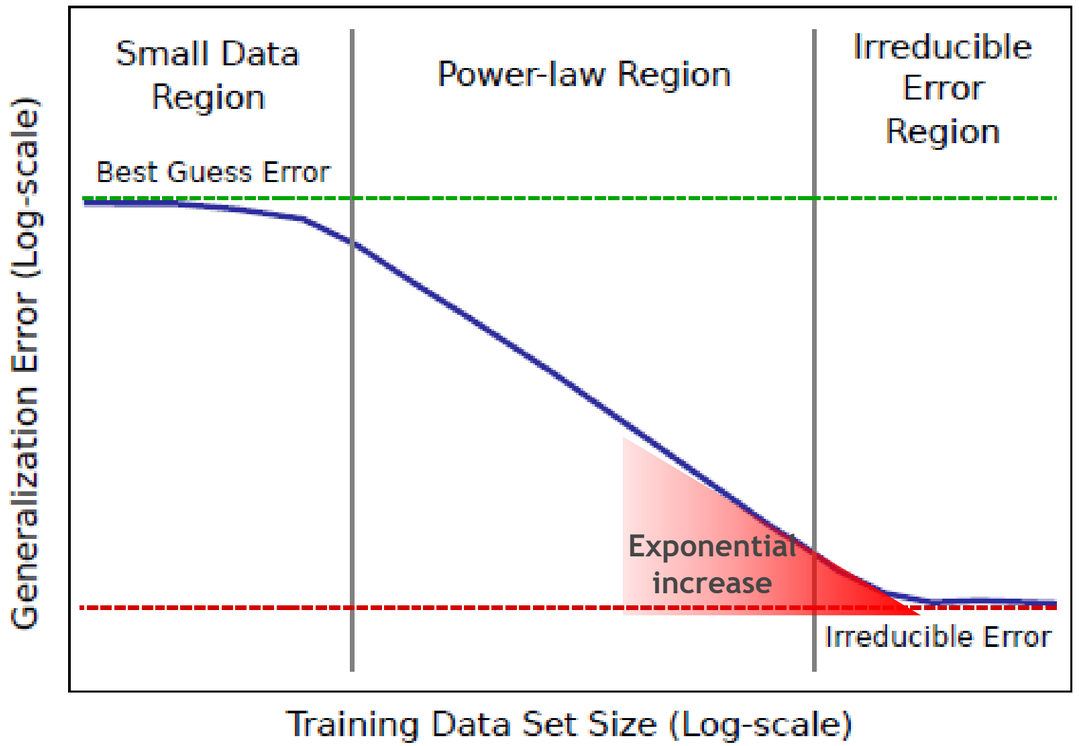


https://developer.nvidia.com/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model


WHY?

THE SCALING LAWS

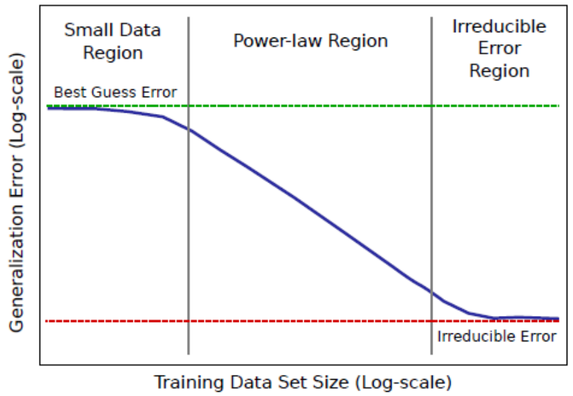
Performance of neural networks increases with model/dataset size

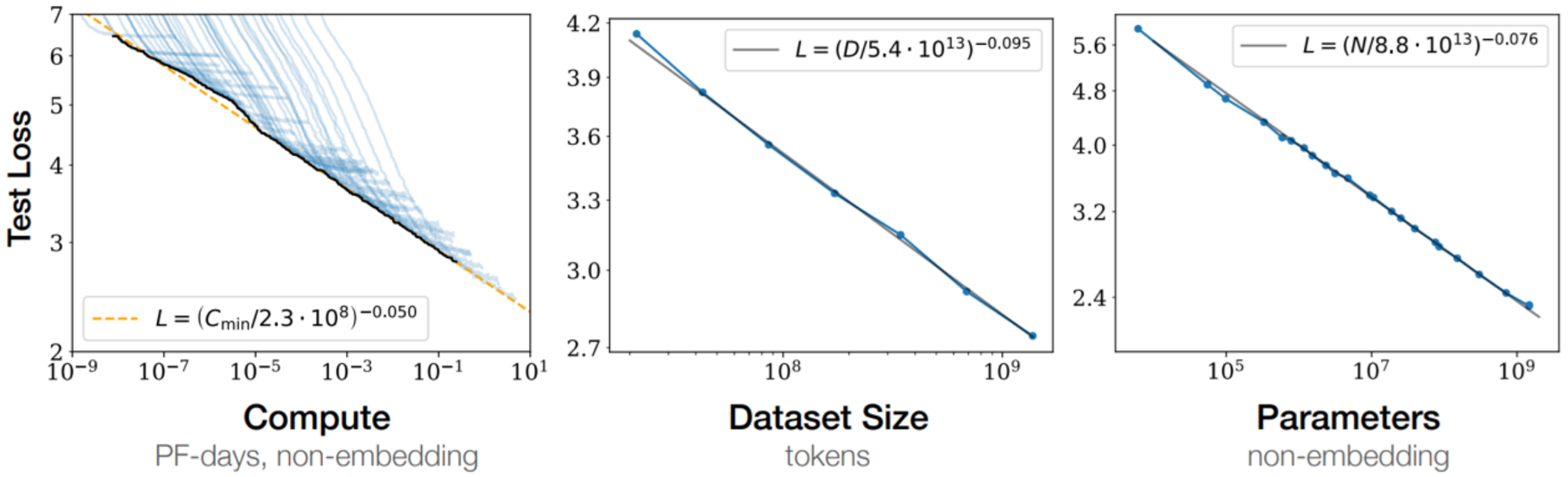


Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., ... & Zhou, Y. (2017). **Deep Learning Scaling is Predictable, Empirically**. arXiv:1712.00409.

EMPIRICAL EVIDENCE

The Scaling Laws in NLP

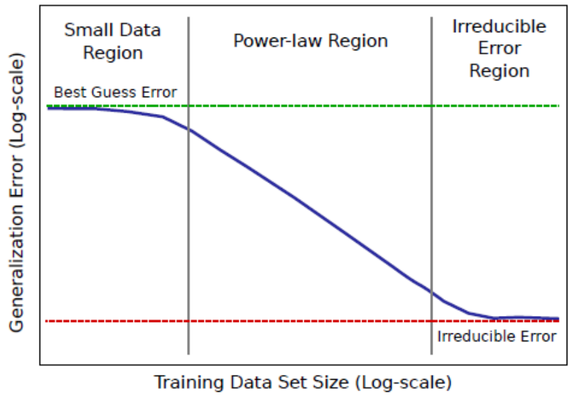


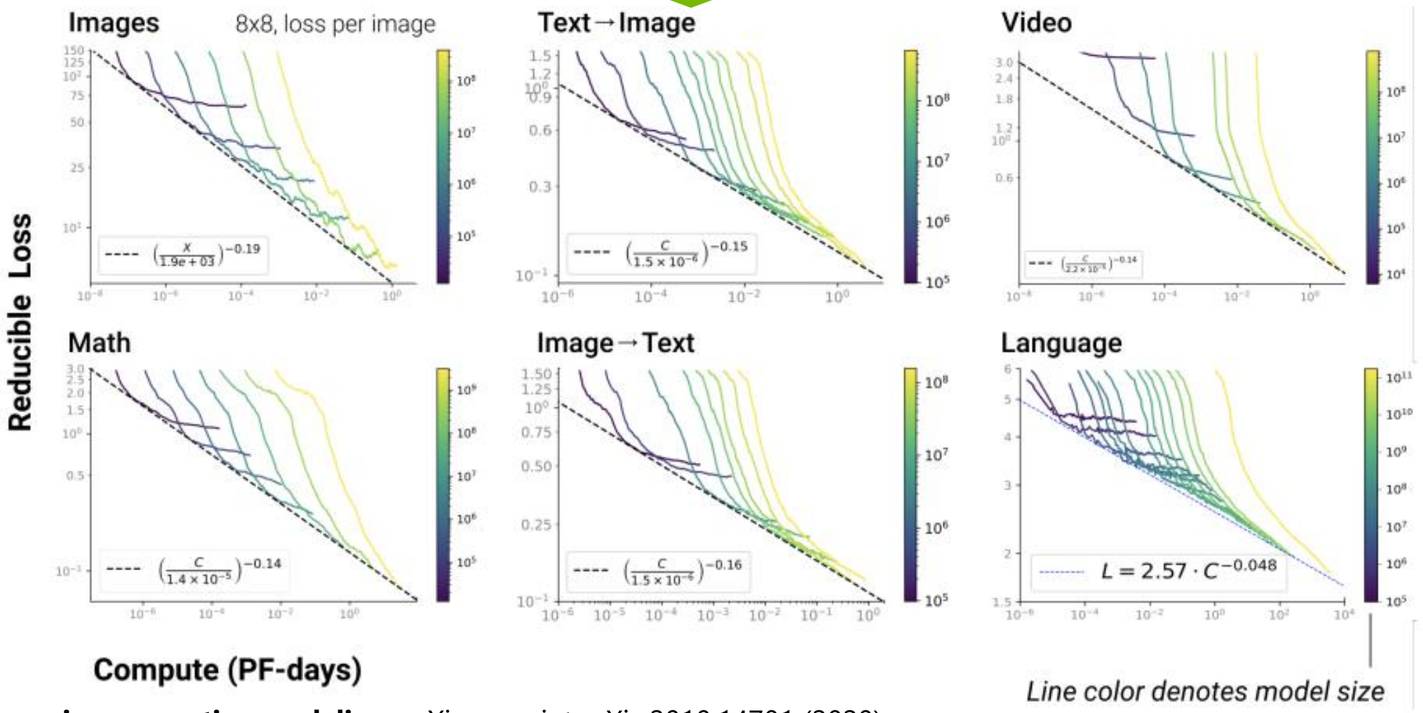


Henighan, Tom, et al. **Scaling laws for autoregressive generative modeling.** arXiv preprint arXiv:2010.14701 (2020).

EMPIRICAL EVIDENCE

The Scaling Laws for Generative models

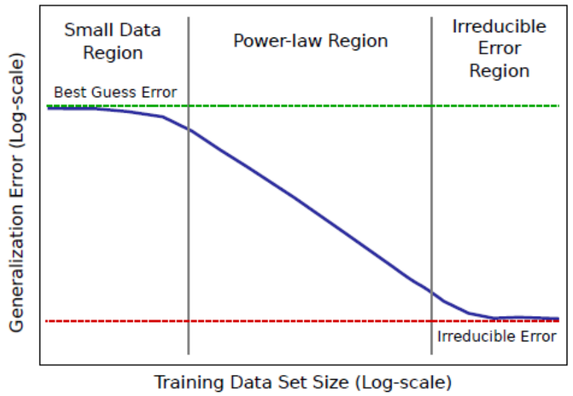


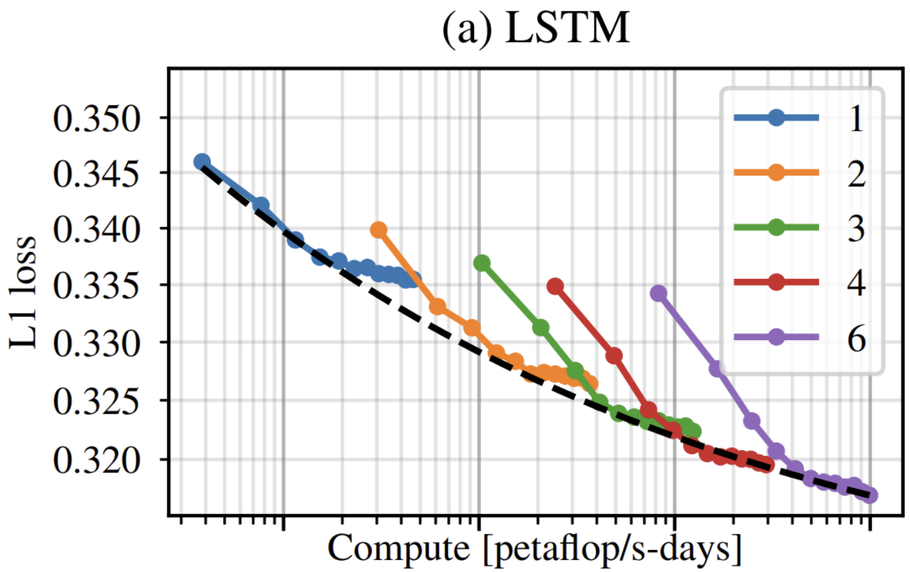


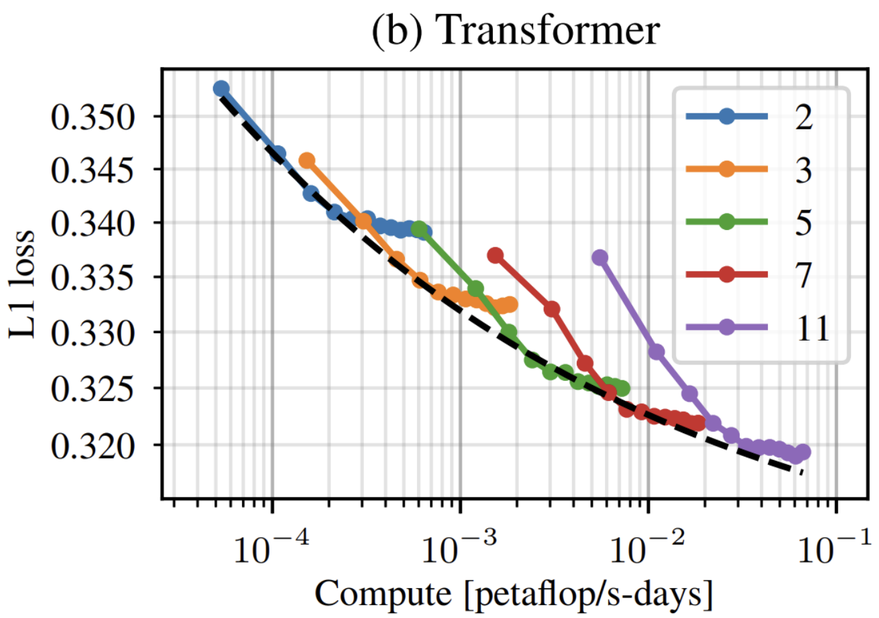
Henighan, Tom, et al. **Scaling laws for autoregressive generative modeling.** arXiv preprint arXiv:2010.14701 (2020).

EMPIRICAL EVIDENCE

The Scaling Laws in Speech



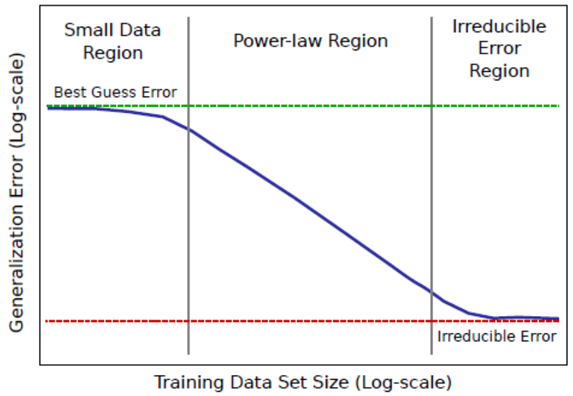


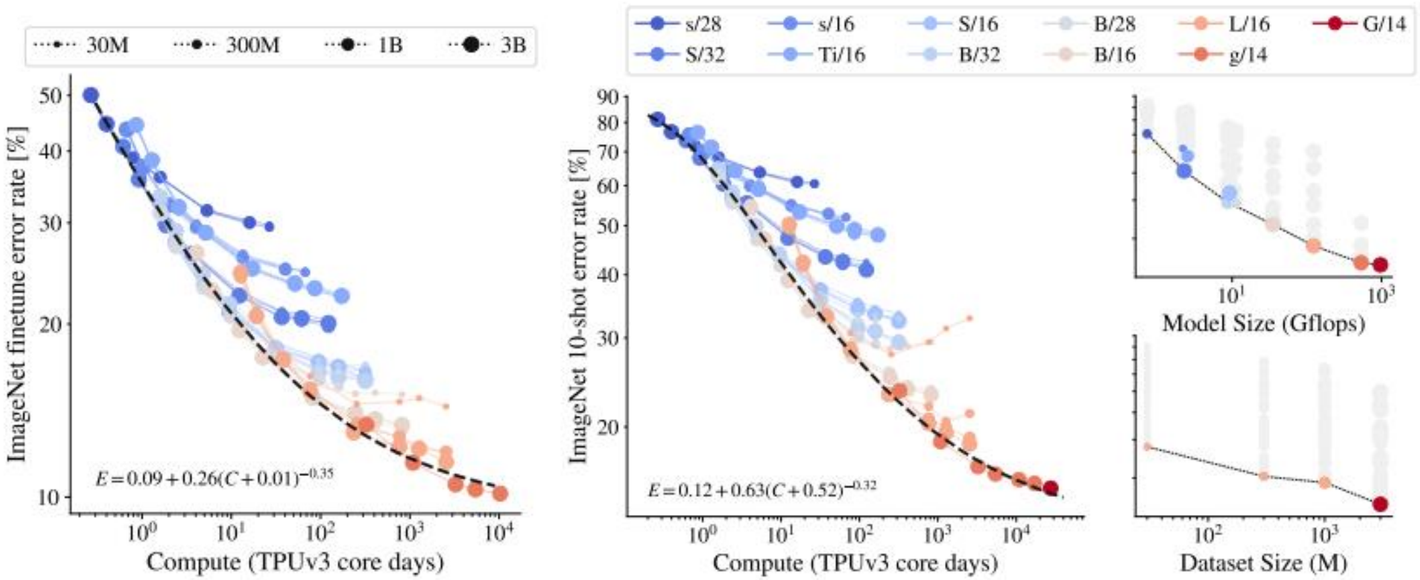


Droppo, Jasha, and Oguz Elibol. **Scaling Laws for Acoustic Models**. arXiv preprint arXiv:2106.09488 (2021).

EMPIRICAL EVIDENCE

The Scaling Laws in Computer Vision



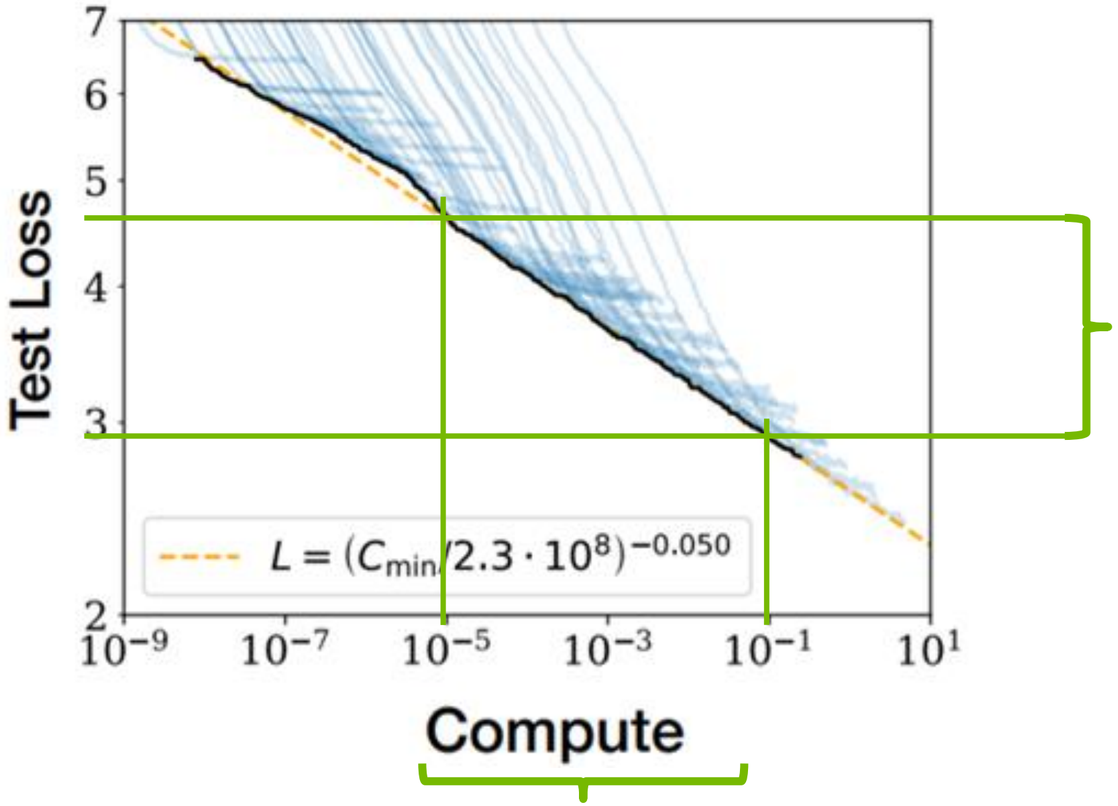


Zhai, Xiaohua, et al. **Scaling vision transformers**. arXiv preprint arXiv:2106.04560 (2021).


BEYOND ACCURACY

ARE LARGE LANGUAGE MODELS WORTH IT?

The cost of incremental improvement



10,000x Increase

Are we building those models only for the small incremental improvement in their performance?

Is it worth all the engineering and computational investment?

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei, Sam McCandlish. **Scaling Laws for Autoregressive Generative Modeling**.2020

FEW SHOT LEARNING

Learning from far fewer examples

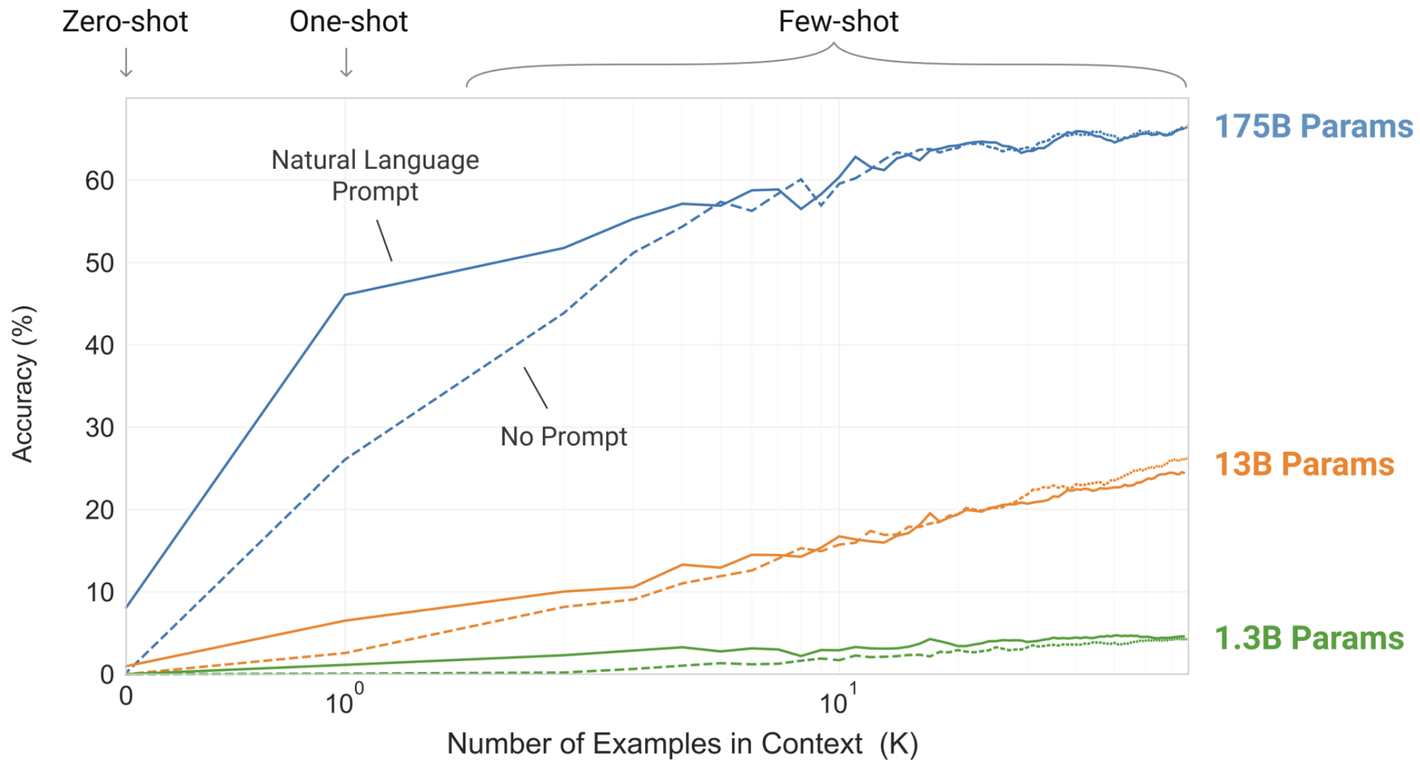
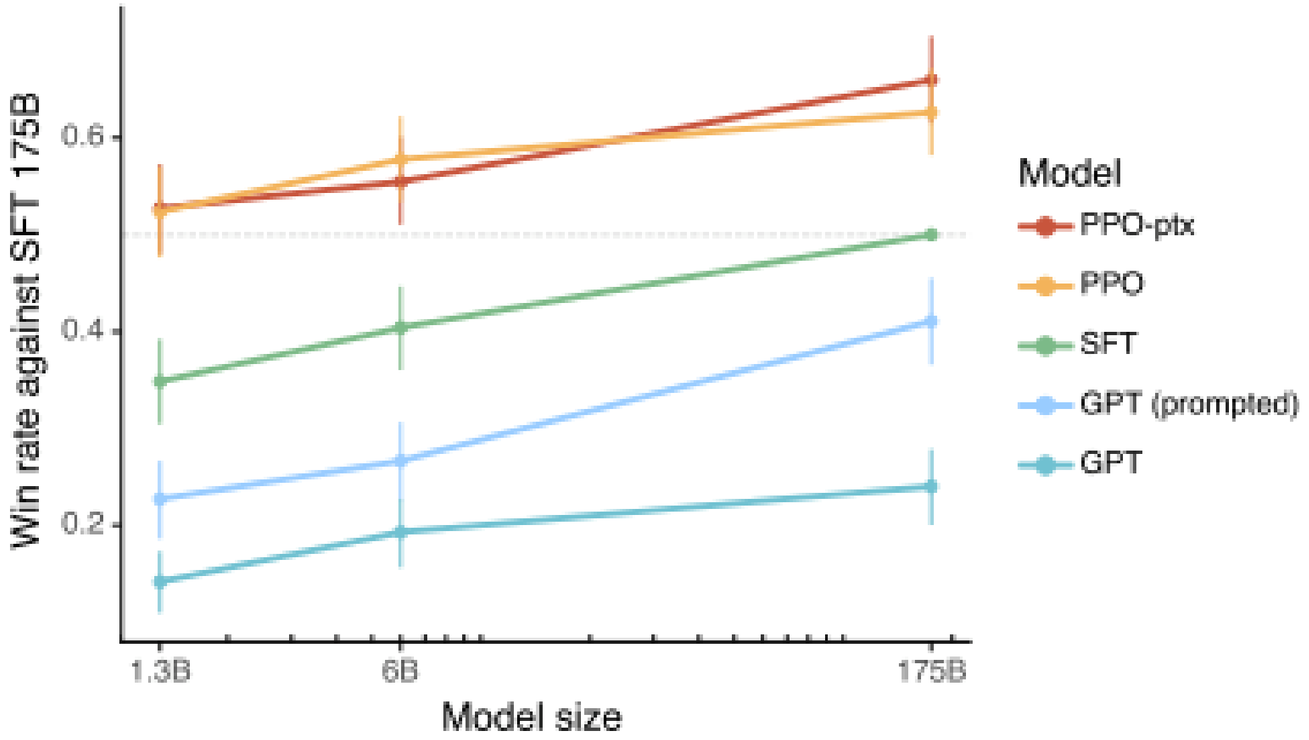


Figure 1.2: Larger models make increasingly efficient use of in-context information. We show in-context learning performance on a simple task requiring the model to remove random symbols from a word, both with and without a natural language task description (see Sec. 3.9.2). The steeper "in-context learning curves" for large models demonstrate improved ability to learn a task from contextual information. We see qualitatively similar behavior across a wide range of tasks.

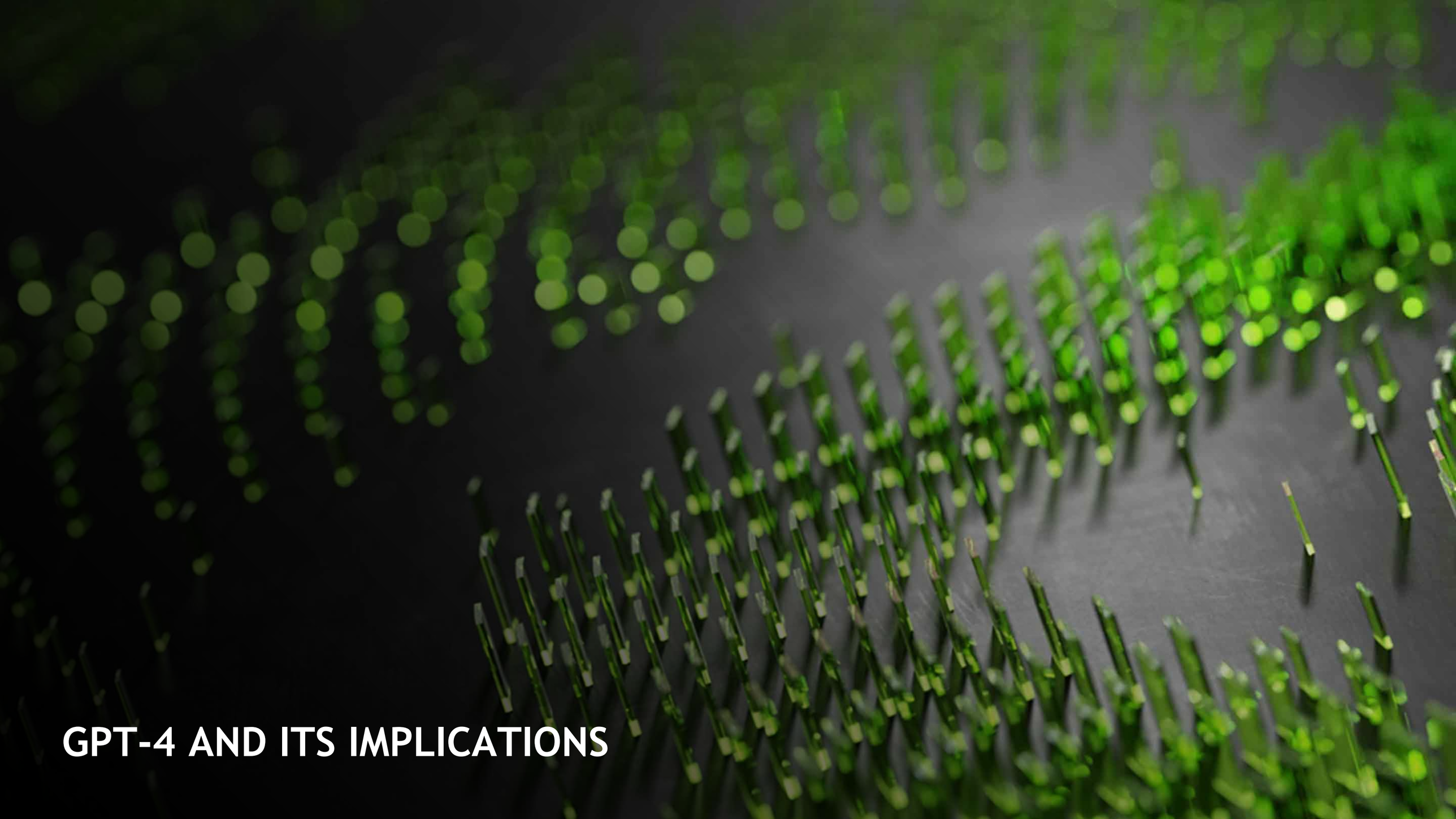
Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Agarwal, S. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*..

FINETUNED LANGUAGE MODELS ARE ZERO SHOT LEARNERS

Exceptional zero shot learning capability



Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730-27744.


GPT-4 AND ITS IMPLICATIONS

Unbelievable Rate of Progress

Major shift in capabilities

Model	**GPT-4**	`text-davinci-003`	Codex(`code-davinci-002`)	CODEGEN-16B
Accuracy	**82%**	65%	39%	30%

Table 1: Zero-shot pass@1 accuracy comparison of different models on HumanEval

Beyond Incremental Improvement to NLP

Exceptional zero shot learning capability

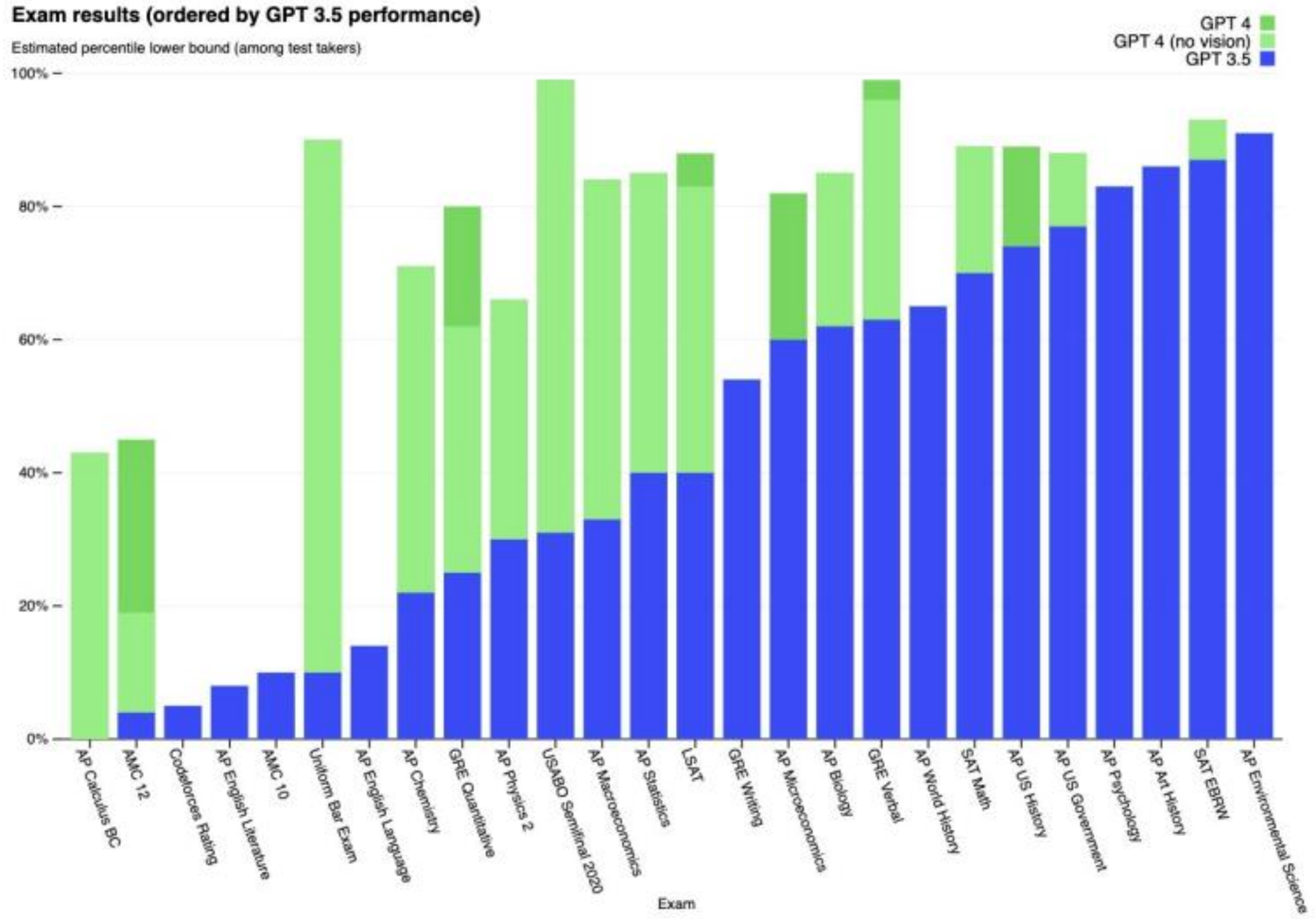


Figure 1: To get a sense of how quickly model capabilities are progressing – consider the jump in exam performance between GPT-3.5 and GPT-4 (OpenAI, 2023b).

Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research

GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models

Tyna Eloundou[1], Sam Manning[1,2], Pamela Mishkin*[1], and Daniel Rock[3]

[1]OpenAI
[2]OpenResearch
[3]University of Pennsylvania

March 27, 2023

Abstract

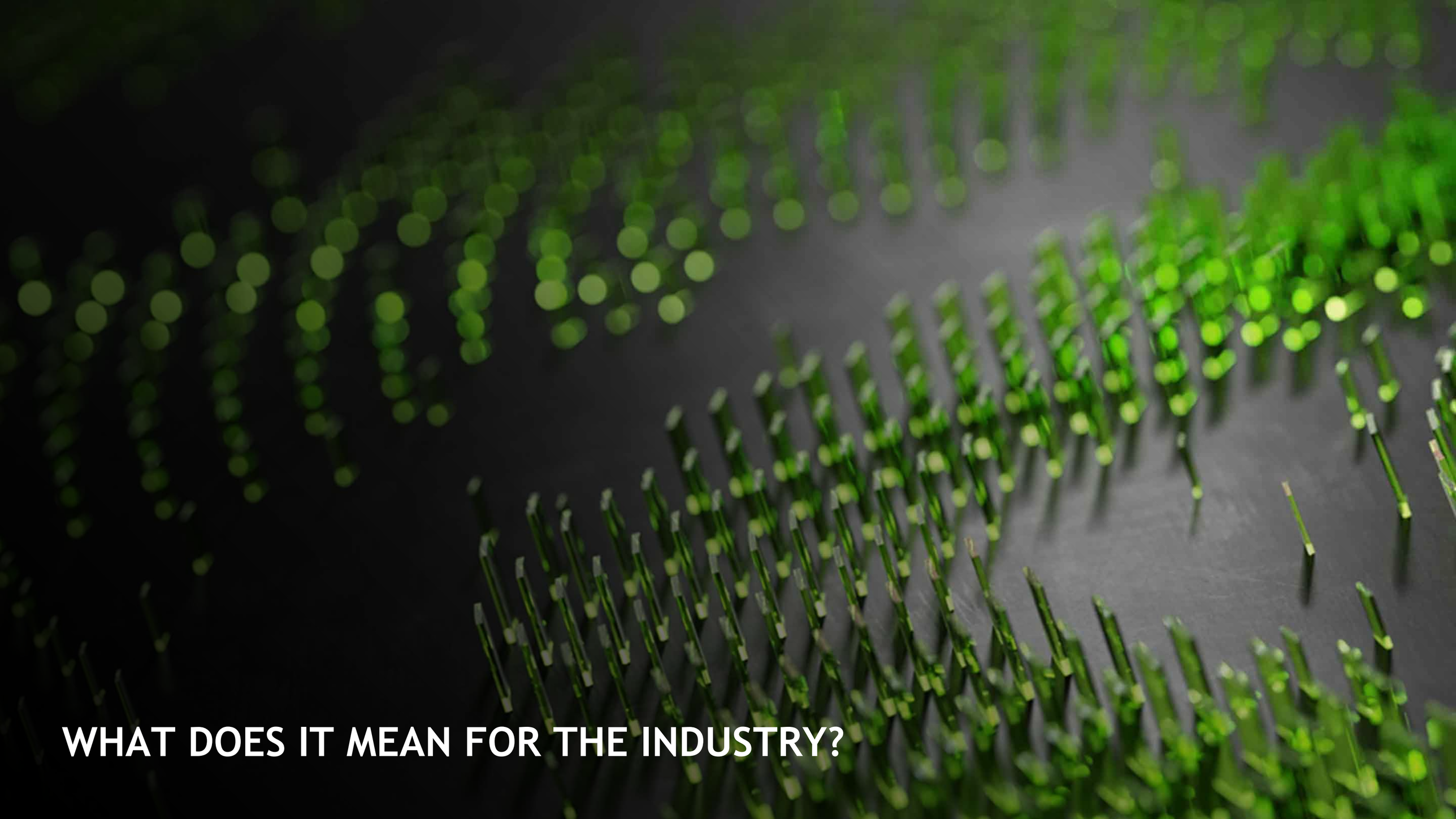
We investigate the potential implications of large language models (LLMs), such as Generative Pre-trained Transformers (GPTs), on the U.S. labor market, focusing on the increased capabilities arising from LLM-powered software compared to LLMs on their own. Using a new rubric, we assess occupations based on their alignment with LLM capabilities, integrating both human expertise and GPT-4 classifications. Our findings reveal that around 80% of the U.S. workforce could have at least 10% of their work tasks affected by the introduction of LLMs, while approximately 19% of workers may see at least 50% of their tasks impacted. We do not make predictions about the development or adoption timeline of such LLMs. The projected effects span all wage levels, with higher-income jobs potentially facing greater exposure to LLM capabilities and LLM-powered software. Significantly, these impacts are not restricted to industries with higher recent productivity growth. Our analysis suggests that, with access to an LLM, about 15% of all worker tasks in the US could be completed significantly faster at the same level of quality. When incorporating software and tooling built on top of LLMs, this share increases to between 47 and 56% of all tasks. This finding implies that LLM-powered software will have a substantial effect on scaling the economic impacts of the underlying models. We conclude that LLMs such as GPTs exhibit traits of general-purpose technologies, indicating that they could have considerable economic, social, and policy implications.

80% of U.S. workforce…
10% of their work tasks affected

With access to an LLM…
47% and 56% of all work tasks could be completed significantly faster

NVIDIA

Impact



WHAT DOES IT MEAN FOR THE INDUSTRY?

OBVIOUS APPLICATIONS

Changing Competitive Landscape

55.8% faster than the control group

Capgemini Google Apple

accenture Microsoft NVIDIA

I write **50 lines** of code per day

+ GitHub Copilot

I write **100-1000** lines of code per day

+ unit tests + comments + documentation

	Easy		Median		Hard		Overall	
pass@k	$k=1$	$k=5$	$k=1$	$k=5$	$k=1$	$k=5$	$k=1$	$k=5$
GPT-4	**68.2**	**86.4**	**40.0**	**60.0**	**10.7**	**14.3**	**38.0**	**53.0**
`text-davinci-003`	50.0	81.8	16.0	34.0	0.0	3.6	19.0	36.0
Codex (`code-davinci-002`)	27.3	50.0	12.0	22.0	3.6	3.6	13.0	23.0
Human (LeetCode users)	72.2		37.7		7.0		38.2	

Table 2: Zero-shot pass@1 and pass@5 accuracies (%) on LeetCode.

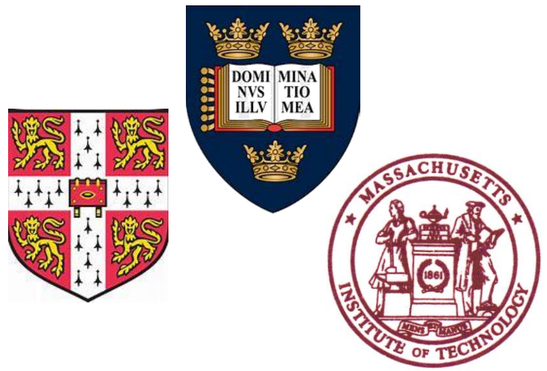
Peng, S., Kalliamvakou, E., Cihon, P., & Demirer, M. (2023). **The impact of ai on developer productivity: Evidence from github copilot**. *arXiv preprint arXiv:2302.06590*.
Sébastien B *et all* (2023). **Sparks of Artificial General Intelligence: Early experiments with GPT-4**. *arXiv:2303.12712*

NVIDIA

Changing Competitive Landscape

55.8% faster than the control group

~~Capgemini~~ ~~Google~~ ~~Apple~~ ~~accenture~~ ~~Microsoft~~ ~~NVIDIA~~



I write **50 lines** of code per day

\+ GitHub Copilot

I write **100-1000** lines of code per day

+ unit tests + comments + documentation

	Easy		Median		Hard		Overall	
pass@k	$k=1$	$k=5$	$k=1$	$k=5$	$k=1$	$k=5$	$k=1$	$k=5$
GPT-4	**68.2**	**86.4**	**40.0**	**60.0**	**10.7**	**14.3**	**38.0**	**53.0**
`text-davinci-003`	50.0	81.8	16.0	34.0	0.0	3.6	19.0	36.0
Codex (`code-davinci-002`)	27.3	50.0	12.0	22.0	3.6	3.6	13.0	23.0
Human (LeetCode users)	72.2		37.7		7.0		38.2	

Table 2: Zero-shot pass@1 and pass@5 accuracies (%) on LeetCode.

Peng, S., Kalliamvakou, E., Cihon, P., & Demirer, M. (2023). **The impact of ai on developer productivity: Evidence from github copilot**. *arXiv preprint arXiv:2302.06590*.
Sébastien B *et all* (2023). **Sparks of Artificial General Intelligence: Early experiments with GPT-4**. *arXiv:2303.12712*

Changing Competitive Landscape

Across countless disciplines

I understand that you are looking for a general plan for processing seismic data using the Madagascar package. Here's an outline of the processing steps, along with the corresponding Madagascar commands:

1. Load the seismic data (e.g., data.su file):
   sfsuread < data.su endian=n > data.rsf
2. Display a few shots from the data:
   sfwindow < data.rsf n1=2 | sfgrey | sfpen
3. Apply trace header manipulation if needed (e.g., set the correct geometry):
   sfput < data.rsf o1=0 d1=1 o2=0 d2=1 > data_geom.rsf
4. Remove noise from the data using a denoising algorithm (e.g., median filter):
   sfmedian < data_geom.rsf rect1=5 rect2=5 > data_denoised.rsf
5. Apply bandpass filtering to emphasize th
   sfbp < data_denoised.rsf flo=5 fhi=60 > d
6. Perform velocity analysis to estimate inte
   sfveltran < data_filtered.rsf semblance=y
7. Create a smoothed velocity model (using
   sfgsmooth < velocity.rsf rect1=20 rect2=2
8. Perform normal moveout (NMO) correctic
   sfnmo < data_filtered.rsf vnmo=velocity_s
9. Stack the NMO-corrected data to create a
   sfstack < data_nmo.rsf > zero_offset.rsf
10. Apply post-stack migration to the zero-of
    sfkirchhoffmig2 < zero_offset.rsf vel=velo
11. Visualize the final migrated section:
    sfgrey < migrated_section.rsf | sf

demo_madagascar.py > ...

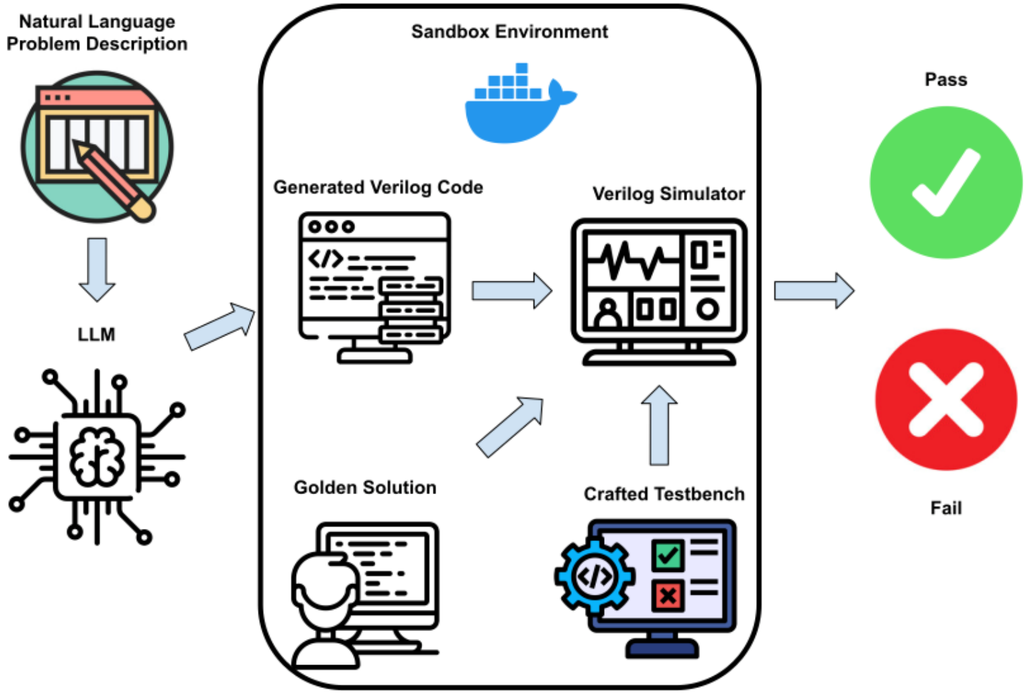
```python
"""https://www.youtube.com/watch?v=r951ZhHrdp0&ab_channel=DataIndependent"""

from langchain.llms.openai import OpenAI
from langchain.agents import load_tools
from langchain.agents import initialize_agent
from langchain.agents import AgentType


if __name__ == "__main__":

    llm = OpenAI(temperature=0, max_tokens=1000)
    tools = load_tools(["terminal", "llm-math"], llm=llm)
    agent = initialize_agent(tools, llm, agent=AgentType.ZERO_SHOT_REACT_DESCRIPTION, verbose=True)


    agent.run("""You have access to the Unix terminal.
                 You will call command line tools from Madagascar software package
                 for seismic data processing.
                 Use "<" to pass input files to the Madagascar funciton.

                 Create a spike sampled with 0.004 sec interval. It should be 1000 samples lon
                 centered on 300, entitled "Welcome". Put it into "spike.rsf".
                 Compute sampling frequency as an inverse of sampling interval.
                 Use it to compute the Nyquist frequency of the signal and bandpass the spike
                 Save the output in "spike_nyquist.rsf".
                 Explain how you computed the Nyquist frequency.
                 Save signal bandpassed below 2 Hz in file "spike_bandpassed.rsf".
                 Use ">" to create the output files.
                 """)
```

```
> Finished chain.
(base) oovcharenko@nvdxb-musk:~/work/aramco/chatbot/langchain/demos$ python d
emo_madagascar.py


> Entering new AgentExecutor chain...
 I need to create a spike, compute the sampling frequency, compute the Nyquis
t frequency, and bandpass the spike below it.
Action: Terminal
Action Input: sfspike n1=1000 d1=0.004 k1=300 label1=Welcome > spike.rsf
Observation:
Thought: I need to compute the sampling frequency
Action: Calculator
Action Input: 1/0.004
Observation: Answer: 250.0

Thought: I need to compute the Nyquist frequency
Action: Calculator
Action Input: 250/2
Observation: Answer: 125.0

Thought: I need to bandpass the spike below the Nyquist frequency
Action: Terminal
Action Input: sfbandpass < spike.rsf fhi=2 > spike_bandpassed.rsf
Observation:
Thought: I now know the final answer
Final Answer: I created a spike sampled with 0.004 sec interval, 1000 samples
 long and centered on 300, entitled "Welcome" and saved it in "spike.rsf". I
computed the sampling frequency as an inverse of sampling interval (1/0.004 =
 250.0) and used it to compute the Nyquist frequency of the signal (250/2 = 1
25.0). I then bandpassed the spike below the Nyquist frequency (fhi=2) and sa
ved the output in "spike_bandpassed.rsf".

> Finished chain.
(base) oovcharenko@nvdxb-musk:~/work/aramco/chatbot/langchain/demos$ sfin < s
pike_bandpass.rsf
in:
    in="/var/tmp/spike_bandpass.rsf@"
    esize=4 type=float form=native
    n1=1000         d1=0.004        o1=0            label1="Time" unit1="s"
        1000 elements 4000 bytes
```

Changing Competitive Landscape

Across countless disciplines

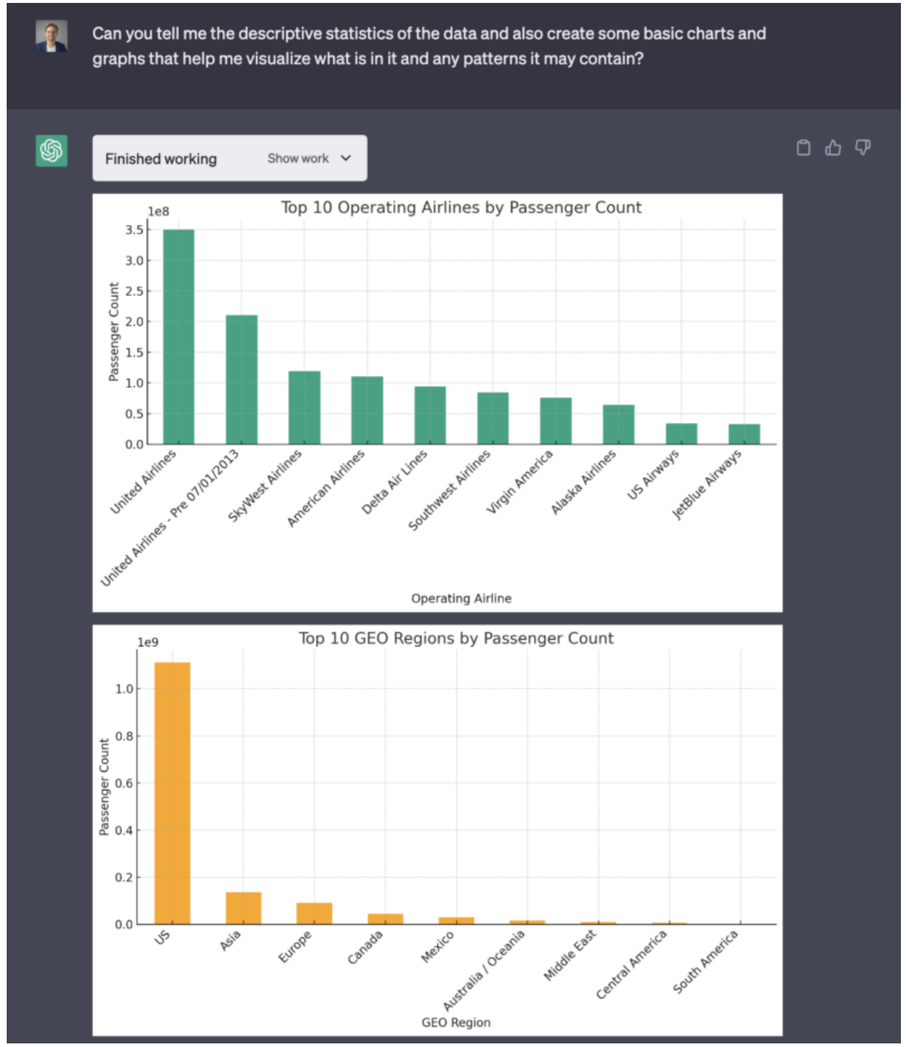
VerilogEval: Evaluating Large Language Models for Verilog Code Generation

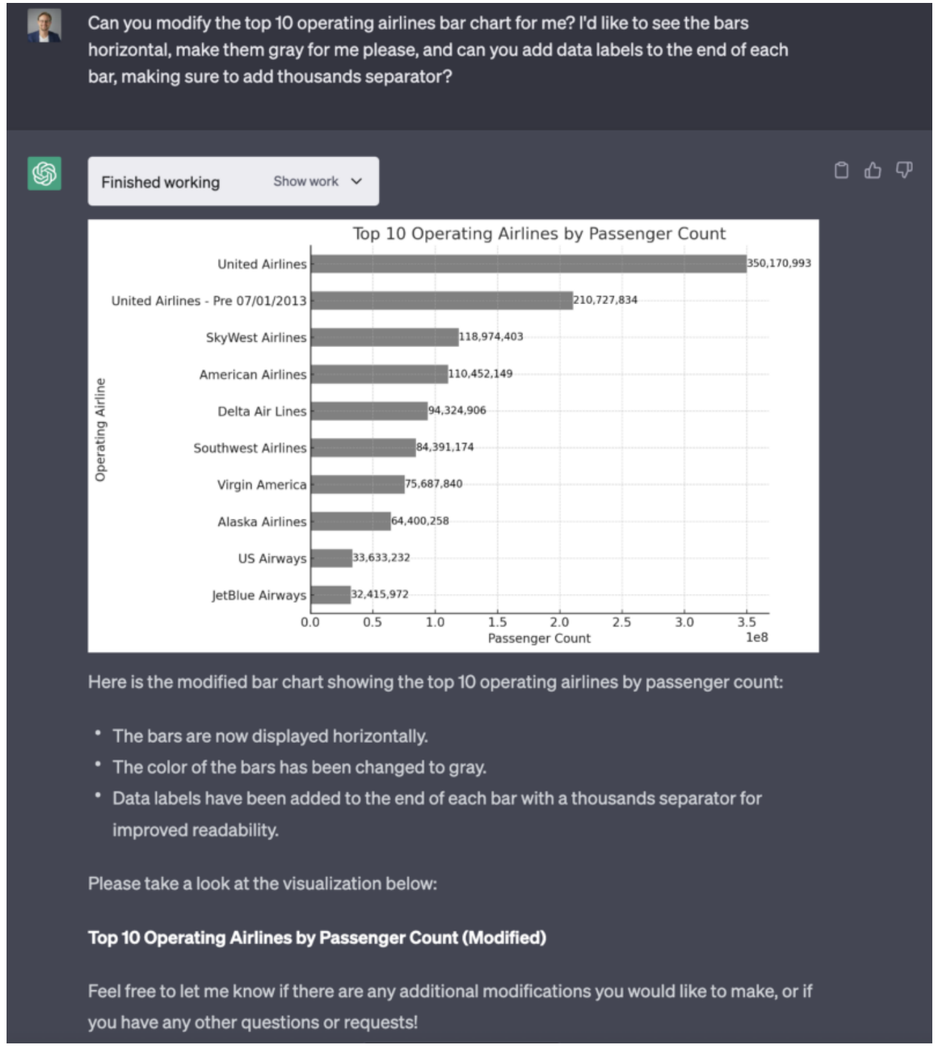


The increasing popularity of large language models (LLMs) has paved the way for their application in diverse domains. This paper proposes a benchmarking framework tailored specifically for evaluating LLM performance in the context of Verilog code generation for hardware design and verification. We present a comprehensive evaluation dataset consisting of 156 problems from the Verilog instructional website HDLBits. The evaluation set consists of a diverse set of Verilog code generation tasks, ranging from simple combinational circuits to complex finite state machines. The Verilog code completions can be automatically tested for functional correctness by comparing the transient simulation outputs of the generated design with a golden solution. We also demonstrate that the Verilog code generation capability of pretrained language models could be improved with supervised fine-tuning by bootstrapping with LLM generated synthetic problem-code pairs.

Changing Competitive Landscape

Reducing barrier to programming





https://dataliteracy.com/code-interpreter-for-chatgpt/

Changing Competitive Landscape

Reducing barrier to programming

https://github.com/KillianLucas/open-interpreter

Changing the way we search

I asked the search engine:

Phind

ANSWER

To create a shape of a tree, we can use the concept of printing patterns in programming. There are multiple ways to create a pattern of a tree, including using loops, recursion, and conditional statements. Let's explore some of the patterns we can use to create a tree shape.

One approach to create a tree shape is to print a pyramid pattern with a trunk. We can use the code provided in programiz.com to print a half pyramid of alphabets. We can modify the code to print a pyramid pattern of asterisks and add a trunk to the pattern.

Here's how we can modify the code to print a tree shape:

```c
#include <stdio.h>
int main() {
   int i, j, k;
   char input, alphabet = '*';
   printf("Enter the height of the tree: ");
   scanf("%c", &input);
   int height = input - '0';
   for (i = 1; i <= height; ++i) {
      for (j = 1; j <= (height - i); ++j) {
         printf(" ");
      }
      for (k = 1; k <= (2*i - 1); ++k) {
         printf("%c", alphabet);
      }
      printf("\n");
   }
   for (i = 1; i <= (height/3); ++i) {
      for (j = 1; j <= (height-2); ++j) {
         printf(" ");
      }
      printf("%c\n", alphabet);
   }
   return 0;
}
```

Ask a followup question

- Can you explain how the loops in the first approach work?
- Can you provide an example of modifying the tree shape using the recursive approach?
- How would you add more branches to the tree using these approaches?
- Is there a way to make the trunk of the tree dynamic in size in the first approach?
- Are there any other programming concepts that could be used to create a tree shape?

In this code, we first take input from the user for the height of the tree. We then use three nested loops to print the pyramid pattern of asterisks. The first loop is for the height of the tree, the second loop is for printing the spaces before the asterisks, and the third loop is for printing the asterisks.

After printing the pyramid pattern, we use two nested loops to print the trunk of the tree. The first loop is for the height

Google

Questions & answers

Quora
Question
I am writing a program to **create** a binary search **tree** and display sorted data. How can I do that in C **programming** easily give m...
Answer · 10 votes
Here is code for in-order traversal in binary search...
More

Bartleby
Question
Programming Language: C Programming1. Make **a code** of implementation of the Binary **Tree** program and explain each line on how it...
Answer · 0 votes
Step 1 Binary tree in c pro...
More

Chegg
Question
1) Consider the following **tree** that holds Entries of type Integer, String: 1,Z 2,D 3,M 4, W 5, H. 6, P 7, V **Write a code** fragment to **create**...
Answer · 0 votes
Solution: I have written the code in Python Language,...
More

Feedback

GeeksforGeeks
https://www.geeksforgeeks.org › create-a-tree-in-level...
Create a tree in level order
Sep 2, 2022 — **Create a tree** in level order ... Given an array of elements, the task is to insert these elements in level order and construct a **tree**.

https://www.geeksforgeeks.org › introduction-to-binar...
Introduction to Binary Tree - Data Structure and Algorithm ...
Mar 15, 2023 — In compilers, Expression **Trees** are used which is an application of binary **trees**. · Huffman **coding trees** are used in data compression algorithms.

People also ask

- How do you make a tree in coding?
- How to create a tree in Java code?
- How to create a tree in C language?
- How do you create a binary tree in programming?

Feedback

Programiz
https://www.programiz.com › dsa › binary-tree
Binary Tree

Chat GPT -OpenAI

write a programming code that takes a letter can creates a shape of a tree

```c
#include <iostream>
#include <string>

using namespace std;

int main(){
  string letter;

  cout << "Enter a letter: ";
  cin >> letter;

  cout << endl;

  cout << letter << letter << letter << endl;
  cout << letter << "  " << letter << endl;
  cout << letter << endl;
  cout << letter << endl;
  cout << letter << endl;
  cout << letter << endl;
  cout << letter << letter << letter << endl;

  return 0;
}
```

NVIDIA.

Changing the way we collate information

Literature review at scale

nature machine intelligence

Article | Open Access | Published: 01 February 2021

An open source machine learning framework for efficient and transparent systematic reviews

Rens van de Schoot ✉, Jonathan de Bruin, Raoul Schram, Parisa Zahedi, Jan de Boer, Felix Weijdema, Bianca Kramer, Martijn Huijts, Maarten Hoogerwerf, Gerbrich Ferdinands, Albert Harkema, Joukje Willemsen, Yongchao Ma, Qixiang Fang, Sybren Hindriks, Lars Tummers & Daniel L. Oberski

Nature Machine Intelligence **3**, 125–133 (2021) | Cite this article

57k Accesses | **125** Citations | **138** Altmetric | Metrics

A preprint version of the article is available at arXiv.

Abstract

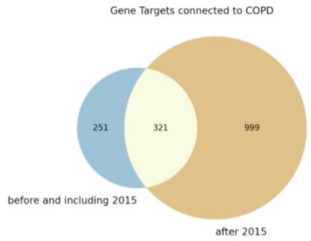
To help researchers conduct a systematic review or meta-analysis as efficiently and transparently as possible, we designed a tool to accelerate the step of screening titles and abstracts. For many tasks—including but not limited to systematic reviews and meta-analyses—the scientific literature needs to be checked systematically. Scholars and practitioners currently screen thousands of studies by hand to determine which studies to include in their review or meta-analysis. This is error prone and inefficient because of extremely imbalanced

https://www.nature.com/articles/s42256-020-00287-7

Changing the way we collate information

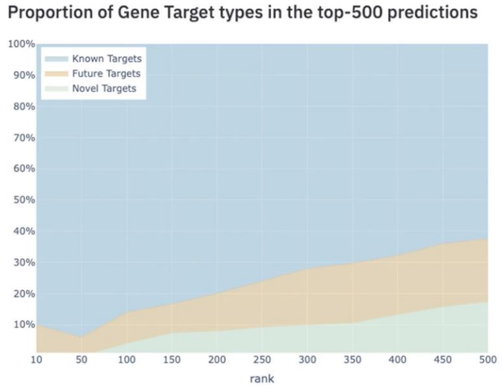
Literature review at scale

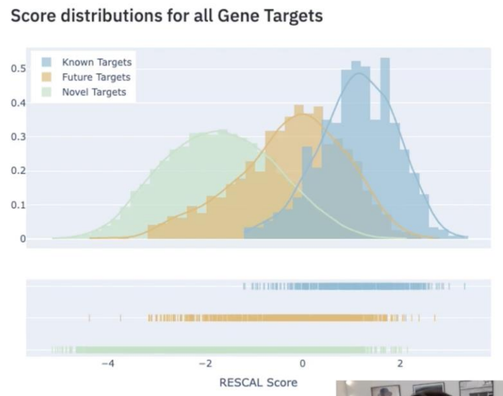
Can we predict novel targets from literature trend?



Pure NLP-based analysis

We can predict novel targets using only NLP





Literature-based model can predict **1 out 6 novel targets before p**

https://www.nvidia.com/en-us/on-demand/session/gtcfall21-a31252/

NVIDIA

Changing the way we manage complex systems

Not just applicable to datacentre infrastructure

Large-language models for automatic cloud incident management

Published May 16, 2023

By Rujia Wang, Principal Research Product Manager; Chetan Bansal, Principal Research Manager; Supriyo GHOSH, Senior Researcher; Tom Zimmermann, Sr. Principal Researcher; Xuchao Zhang, Senior Researcher; Saravan Rajmohan, Partner Director AI and Applied Research

Share this page

This research was accepted by the IEEE/ACM International Conference on Software Engineering (ICSE), which is a forum for researchers, practitioners, and educators to gather, present, and discuss the most recent innovations, trends, experiences, and issues in the field of software engineering.

CHANGING THE WAY WE APPROACH COMPLEX PROBLEMS

LLM in mathematics

Large Language Model for Science: A Study on P vs. NP

Qingxiu Dong [*1,2] Li Dong [*1] Ke Xu [*3]
Guangyan Zhou [4] Yaru Hao [1] Zhifang Sui [2] Furu Wei [1]
https://aka.ms/GeneralAI

Abstract

In this work, we use large language models (LLMs) to augment and accelerate research on the P versus NP problem, one of the most important open problems in theoretical computer science and mathematics. Specifically, we propose Socratic reasoning, a general framework that promotes in-depth thinking with LLMs for complex problem-solving. Socratic reasoning encourages LLMs to recursively discover, solve, and integrate problems while facilitating self-evaluation and refinement. Our pilot study on the P vs. NP problem shows that GPT-4 successfully produces a proof schema and engages in rigorous reasoning throughout 97 dialogue turns, concluding "P $\neq$ NP", which is in alignment with (Xu and Zhou, 2023). The investigation uncovers novel insights within the extensive solution space of LLMs, shedding light on LLM for Science.

MathPrompter: Mathematical Reasoning using Large Language Models

Shima Imani, Liang Du, Harsh Shrivastava
Microsoft Research, Redmond
Contact: shimaimani@microsoft.com

Abstract

Large Language Models (LLMs) have limited performance when solving arithmetic reasoning tasks and often provide incorrect answers. Unlike natural language understanding, math problems typically have a single correct answer, making the task of generating accurate solutions more challenging for LLMs. To the best of our knowledge, we are not aware of any LLMs that indicate their level of confidence in their responses which fuels a trust deficit in these models impeding their adoption. To address this deficiency, we propose 'MathPrompter', a technique that improves performance of LLMs on arithmetic problems along with increased reliance in the predictions. MathPrompter uses the Zero-shot chain-of-thought prompting technique to generate multiple Algebraic expressions or Python functions to solve the same math problem in different ways and thereby raise the confidence level in the output results. This is in contrast to other prompt based CoT methods, where there is no check on the validity of the intermediate steps followed. Our technique improves over state-of-the-art on the MultiArith dataset (78.7% $\rightarrow$ 92.5%) evaluated using 175B parameter GPT-based LLM.

BEYOND GENERIC MODELS

Science

Galactica: A Large Language Model for Science

Ross Taylor **Marcin Kardas** **Guillem Cucurull**

Thomas Scialom **Anthony Hartshorn** **Elvis Saravia**

Andrew Poulton **Viktor Kerkez** **Robert Stojnic**

Meta AI

Abstract

Information overload is a major obstacle to scientific progress. The explosive growth in scientific literature and data has made it ever harder to discover useful insights in a large mass of information. Today scientific knowledge is accessed through search engines, but they are unable to organize scientific knowledge alone. In this paper we introduce Galactica: a large language model that can store, combine and reason about scientific knowledge. We train on a large scientific corpus of papers, reference material, knowledge bases and many other sources. We outperform existing models on a range of scientific tasks. On technical knowledge probes such as LaTeX equations, Galactica outperforms the latest GPT-3 by 68.2% versus 49.0%. Galactica also performs well on reasoning, outperforming Chinchilla on mathematical MMLU by 41.3% to 35.7%, and PaLM 540B on MATH with a score of 20.4% versus 8.8%. It also sets a new state-of-the-art on downstream tasks such as PubMedQA and MedMCQA dev of 77.6% and 52.9%. And despite not being trained on a general corpus, Galactica outperforms BLOOM and OPT-175B on BIG-bench. We believe these results demonstrate the potential for language models as a new interface for science. We open source the model for the benefit of the scientific community[1].

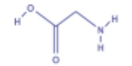
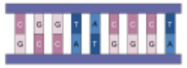
Modality	Entity	Sequence	
Text	Abell 370	`Abell 370 is a cluster...`	
LaTeX	Schwarzschild radius	`r_{s} = \frac{2GM}{c^2}`	$r_s = \frac{2GM}{c^2}$
Code	Transformer	`class Transformer(nn.Module)`	
SMILES	Glycine	`C(C(=O)O)N`	
AA Sequence	Collagen α-1(II) chain	`MIRLGAPQTL..`	
DNA Sequence	Human genome	`CGGTACCCTC..`	

Table 1: Tokenizing Nature. Galactica trains on text sequences that represent scientific phenomena.

BEYOND GENERIC MODELS

Science

A Large Language Model for Electronic Health Records

Authors: Xi Yang[1,2], Aokun Chen[1,2], Nima PourNejatian[3], Hoo Chang Shin[3], Kaleb E Smith[3], Christopher Parisien[3], Colin Compas[3], Cheryl Martin[3], Anthony B Costa[3], Mona G Flores[3], Ying Zhang[4], Tanja Magoc[5], Christopher A Harle[1,5], Gloria Lipori[5,6], Duane A Mitchell[6], William R Hogan[1], Elizabeth A Shenkman[1], Jiang Bian[1,2], Yonghui Wu[1,2 *]

Affiliations:

[1]Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, Florida, USA.

[2]Cancer Informatics and eHealth core, University of Florida Health Cancer Center, Gainesville, Florida, USA.

[3]NVIDIA, Santa Clara, California, USA.

[4]Research Computing, University of Florida, Gainesville, Florida, USA.

[5]Integrated Data Repository Research Services, University of Florida, Gainesville, Florida, USA.

[6]Lillian S. Wells Department of Neurosurgery, UF Clinical and Translational Science Institute, University of Florida.

BloombergGPT: A Large Language Model for Finance

Shijie Wu[1,*], Ozan İrsoy[1,*], Steven Lu[1,*], Vadim Dabravolski[1], Mark Dredze[1,3], Sebastian Gehrmann[1], Prabhanjan Kambadur[1], David Rosenberg[2], Gideon Mann[1]

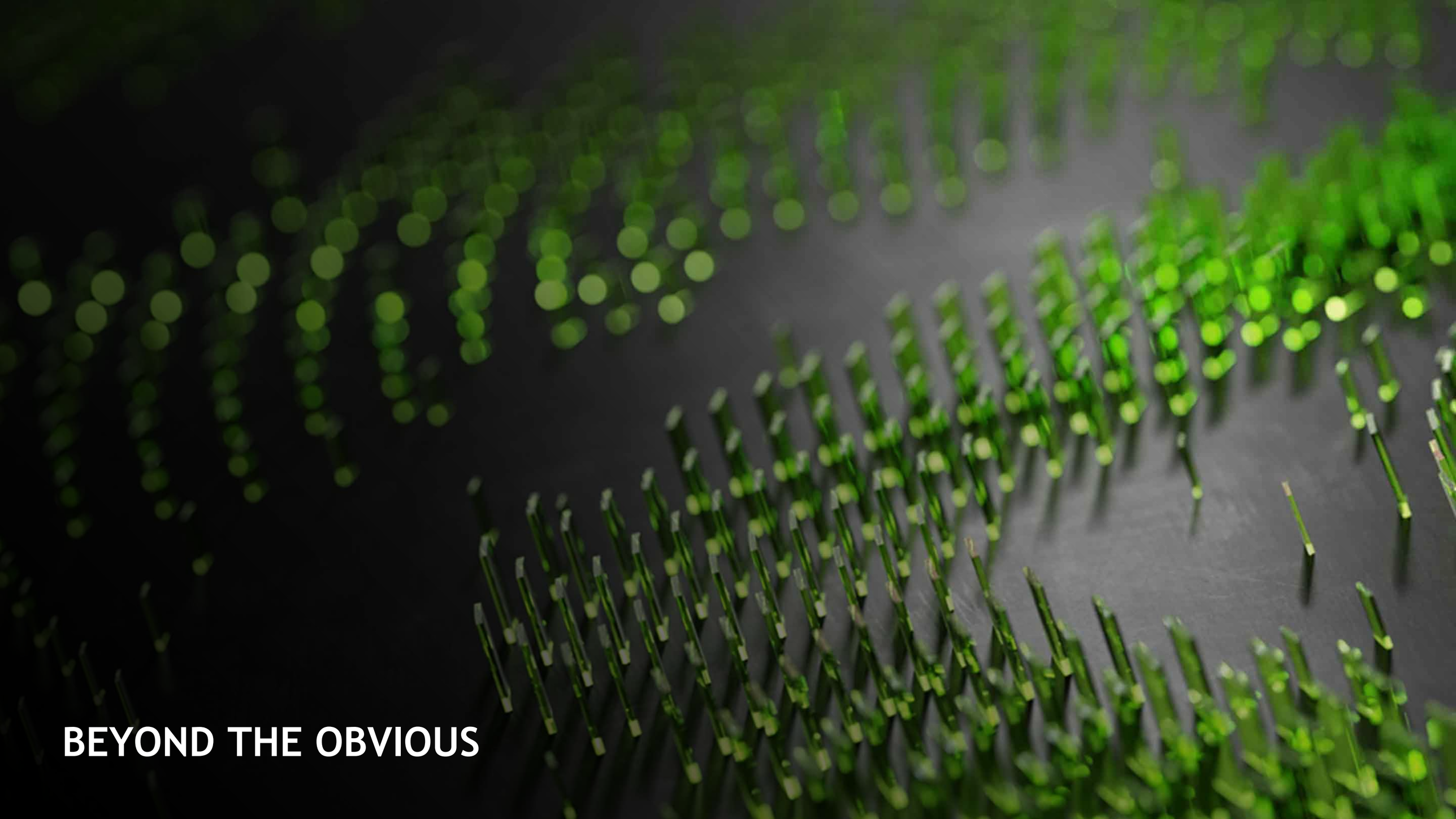
[1] **Bloomberg, New York, NY USA**

[2] **Bloomberg, Toronto, ON Canada**

[3] **Computer Science, Johns Hopkins University, Baltimore, MD USA**

Abstract

The use of NLP in the realm of financial technology is broad and complex, with applications ranging from sentiment analysis and named entity recognition to question answering. Large Language Models (LLMs) have been shown to be effective on a variety of tasks; however, no LLM specialized for the financial domain has been reported in literature. In this work, we present BloombergGPT, a 50 billion parameter language model that is trained on a wide range of financial data. We construct a 363 billion token dataset based on Bloomberg's extensive data sources, perhaps the largest domain-specific dataset yet, augmented with 345 billion tokens from general purpose datasets. We validate BloombergGPT on standard LLM benchmarks, open financial benchmarks, and a suite of internal benchmarks that most accurately reflect our intended usage. Our mixed dataset training leads to a model that outperforms existing models on financial tasks by significant margins without sacrificing performance on general LLM benchmarks. Additionally, we explain our modeling choices, training process, and evaluation methodology. We release Training Chronicles (Appendix C) detailing our experience in training BloombergGPT.


BEYOND THE OBVIOUS

Beyond the Obvious

We can only see the first wave of business models affected

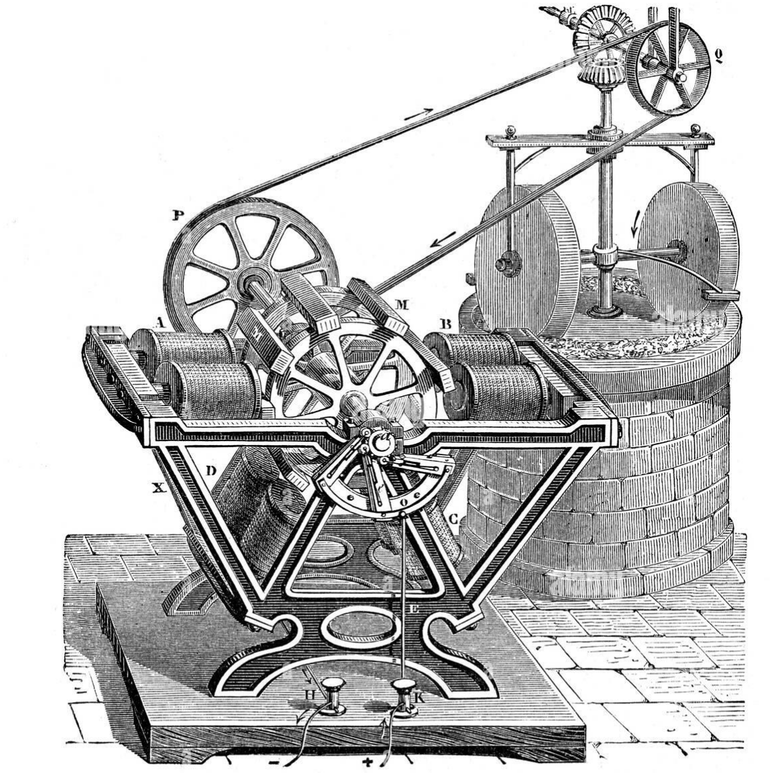
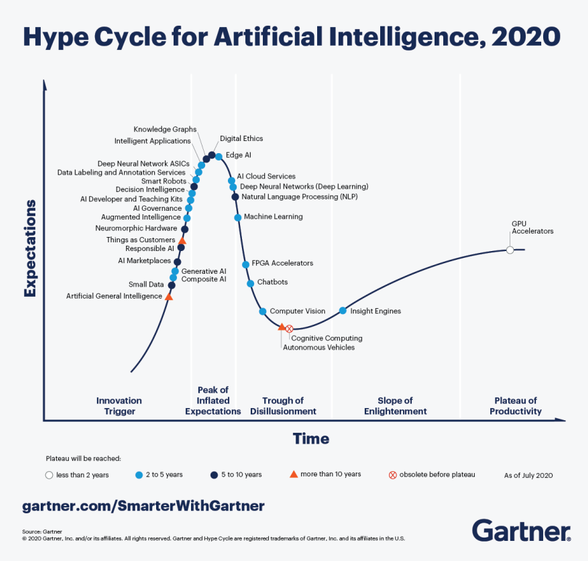
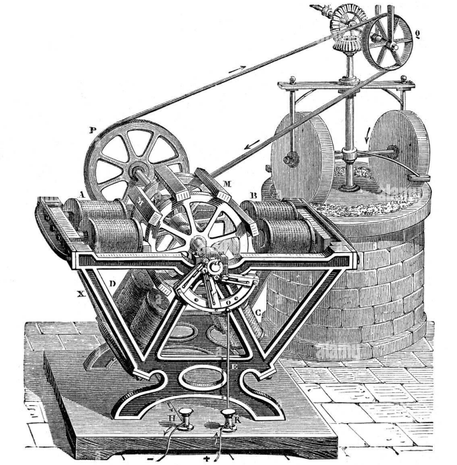


Fig. 314. — Moteur Froment attelé à une paire de meules.

Transforming Impossible into Feasible

Future of books / reports





Transforming Impossible into Feasible

Democratizing access to education

Courses ▾ Search Khan Academy Get AI Guide Donate Log in Sign up

For every student,
every classroom.
Real results.

We're a nonprofit with the mission to provide a free, world-class education for anyone, anywhere.

Learners Teachers Parents

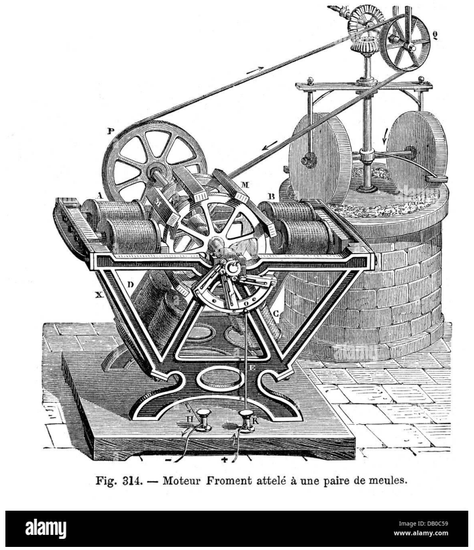


Fig. 314. — Moteur Froment attelé à une paire de meules.

alamy Image ID: D80C59 www.alamy.com

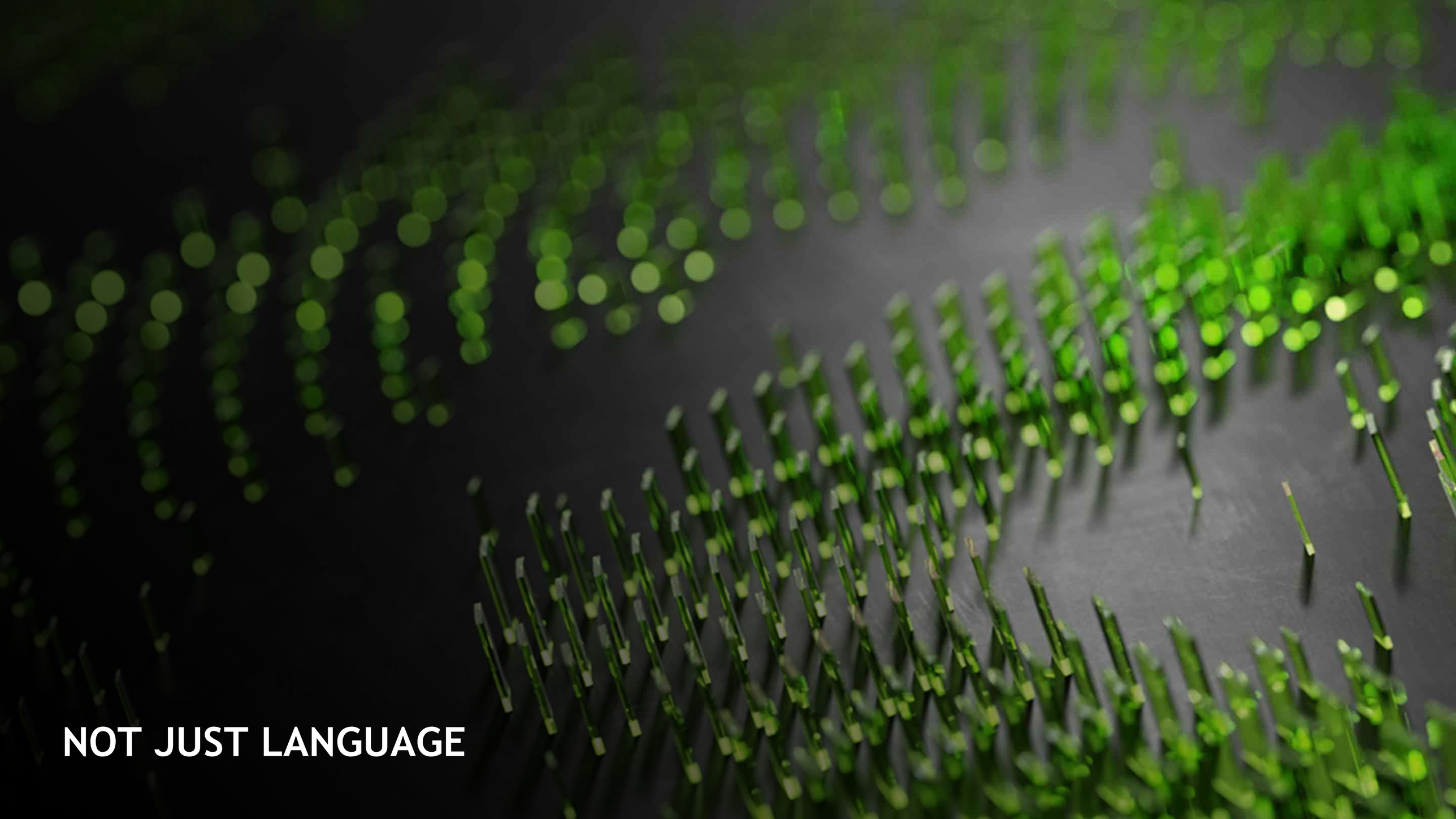
Bill Gates says AI chatbots like ChatGPT can replace human teachers

AI-powered tutors could be a more economical solution for parents who can't afford a human teacher.

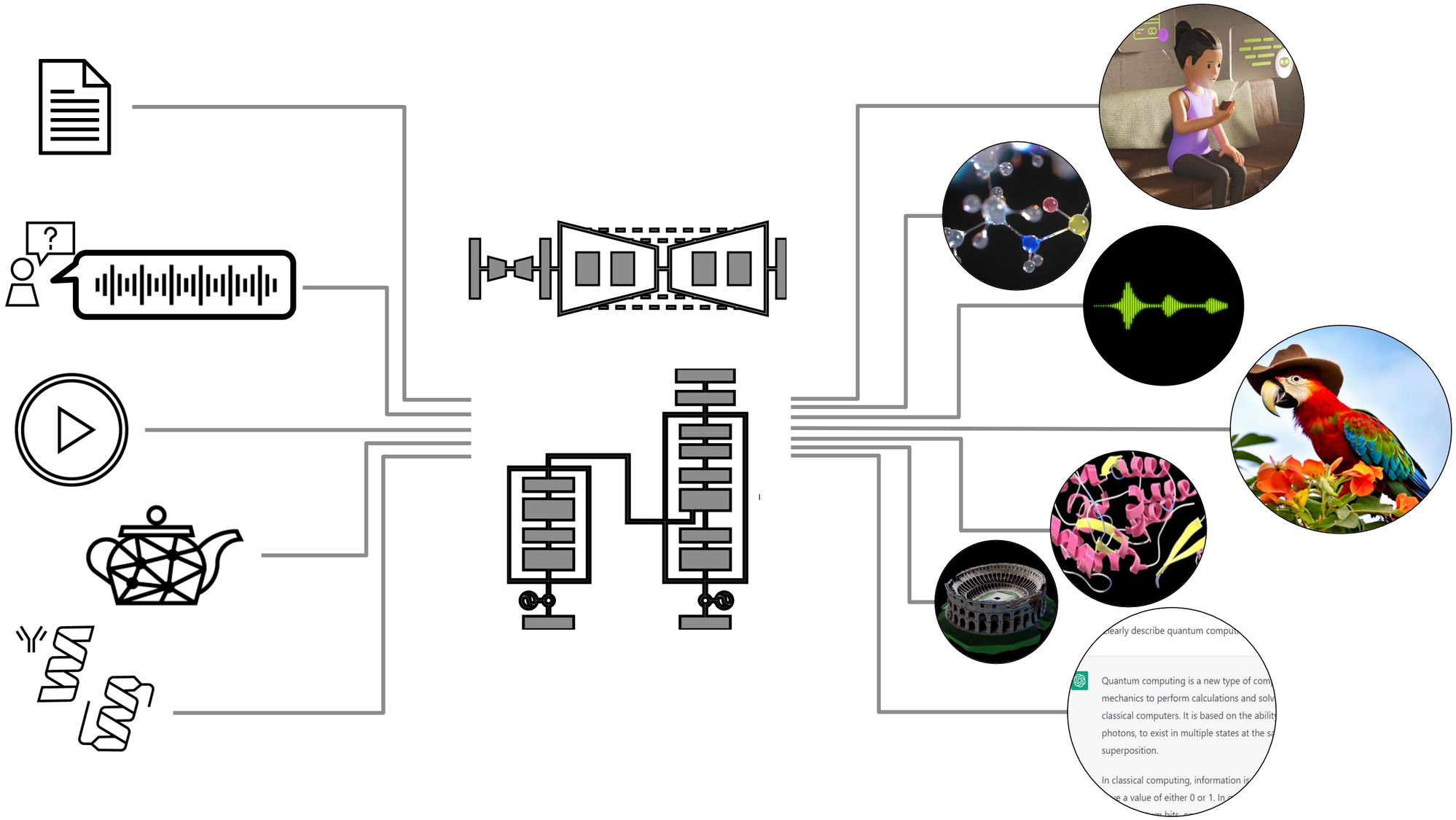
By Vinay Patel @VinayPatelBlogs
04/27/23 AT 7:28 AM BST

Bill Gates beleives AI chatbots will soon replace human teachers. (PHOTO: JOHN LAMPARSKI/GETTY IMAGES)

NVIDIA


NOT JUST LANGUAGE

WHAT IS GENERATIVE AI?



BIOLOGY

Nucleotide transformer

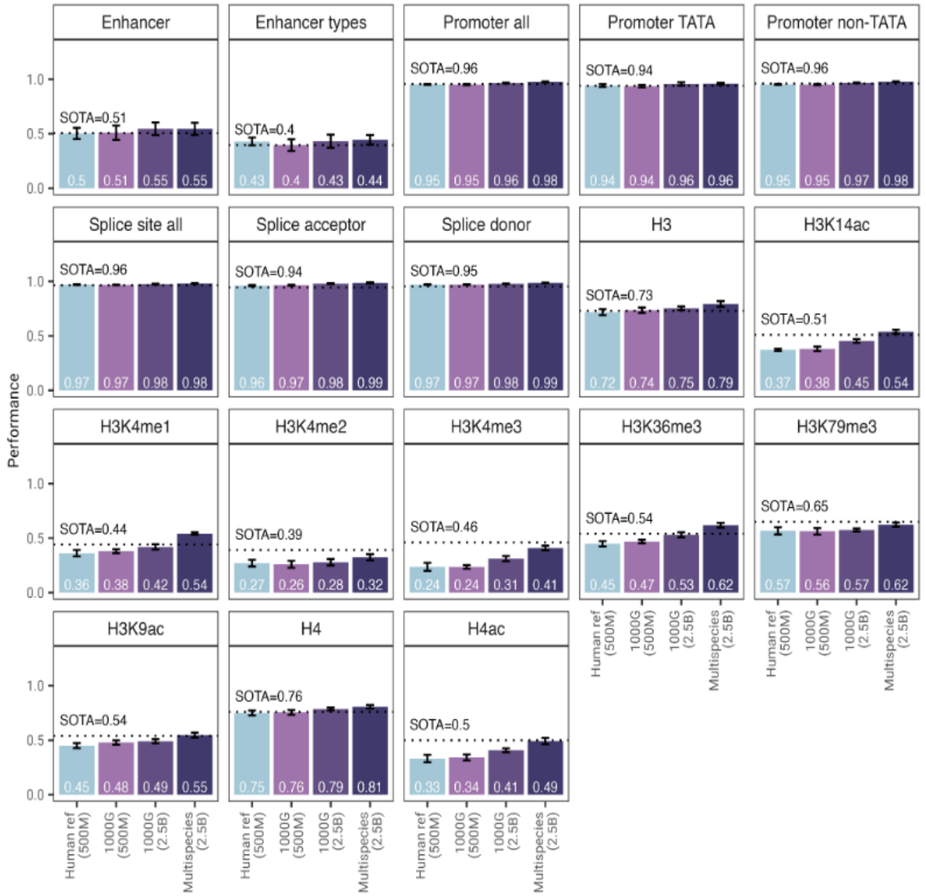


Fig. 1: The Nucleotide Transformer model matches or outperforms 15 out of 18 downstream tasks using fine-tuning. We show the performance results across downstream tasks for fine-tuned transformer models. Error bars represent 2 SDs derived from 10-fold cross-validation. The performance metrics for the state-of-the-art (SOTA) models are shown as horizontal dotted lines.

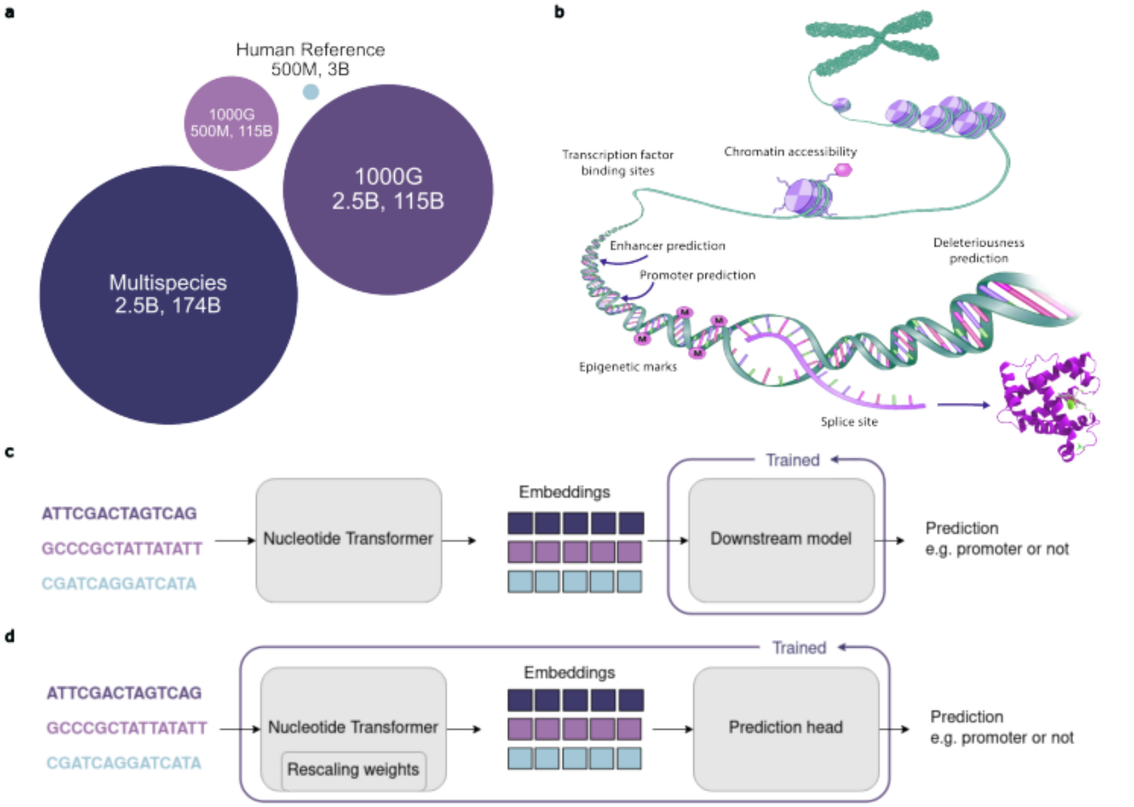
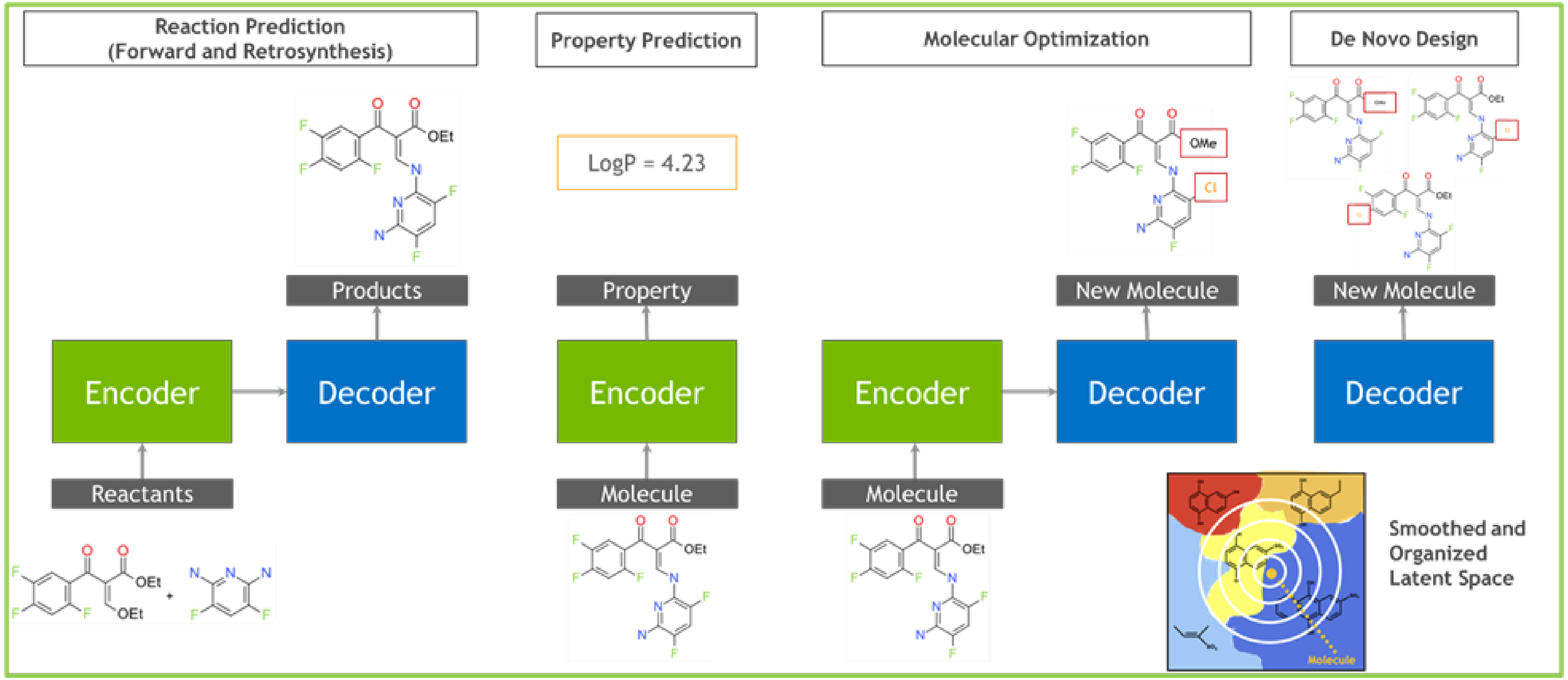


Figure 1: **The Nucleotide Transformer: a Masked Language Model trained for Genomics Prediction**. **a)** Training datasets and parameter sizes of the language models. **b)** Graphical representation of genomic features considered for prediction tasks. **c)** Overview of the Nucleotide Transformer training and application for downstream genomic prediction tasks through probing. **d)** Overview of the Nucleotide Transformer training and application for downstream genomic prediction tasks through fine-tuning.

CHEMISTRY / DRUG DISCOVERY

MegaMolBart



CHEMISTRY / DRUG DISCOVERY

MolGPT

MolGPT: Molecular Generation Using a Transformer-Decoder Model

Viraj Bagal, Rishal Aggarwal, P. K. Vinod, and U. Deva Priyakumar*

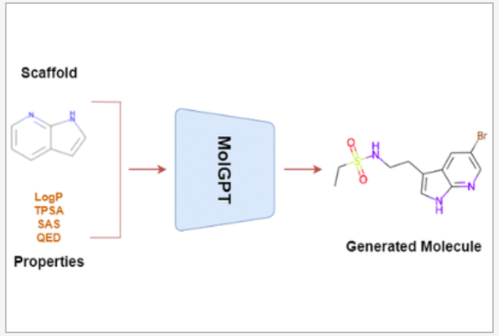
Cite this: *J. Chem. Inf. Model.* 2022, 62, 9, 2064–2076
Publication Date: October 25, 2021
https://doi.org/10.1021/acs.jcim.1c00600

Journal of Chemical Information and Modeling

SUBJECTS: Molecular modeling, Molecular properties, Molecules, Partition coefficient, Scaffolds

Abstract

Application of deep learning techniques for *de novo* generation of molecules, termed as inverse molecular design, has been gaining enormous traction in drug design. The representation of molecules in SMILES notation as a string of characters enables the usage of state of the art models in natural language processing, such as Transformers, for molecular design in general. Inspired by generative pre-training (GPT) models that have been shown to be successful in generating meaningful text, we train a transformer-decoder on the next token prediction task using masked self-attention for the generation of druglike molecules in this study. We show that our model, MolGPT, performs on par with other previously proposed modern machine learning frameworks for molecular generation in terms of generating valid, unique, and novel molecules. Furthermore, we demonstrate that the model can be trained conditionally to control multiple properties of the generated molecules. We also show that the model can be used to generate molecules with desired scaffolds as well as desired molecular properties by conditioning the generation on scaffold SMILES strings of desired scaffolds and property values. Using saliency maps, we highlight the interpretability of the generative process of the model.



BIOLOGY

Beyond academic research

PHARMA.AI

PandaOmics

Discover and Prioritize

Novel Targets

Enabling multi-omics target discovery and deep biology analysis engine to considerably reduce required time

Chemistry42

Generate

Novel Molecules

Find novel lead-like molecules through this automated, machine learning de-novo drug design and scalable engineering platform

InClinico

Design and predict

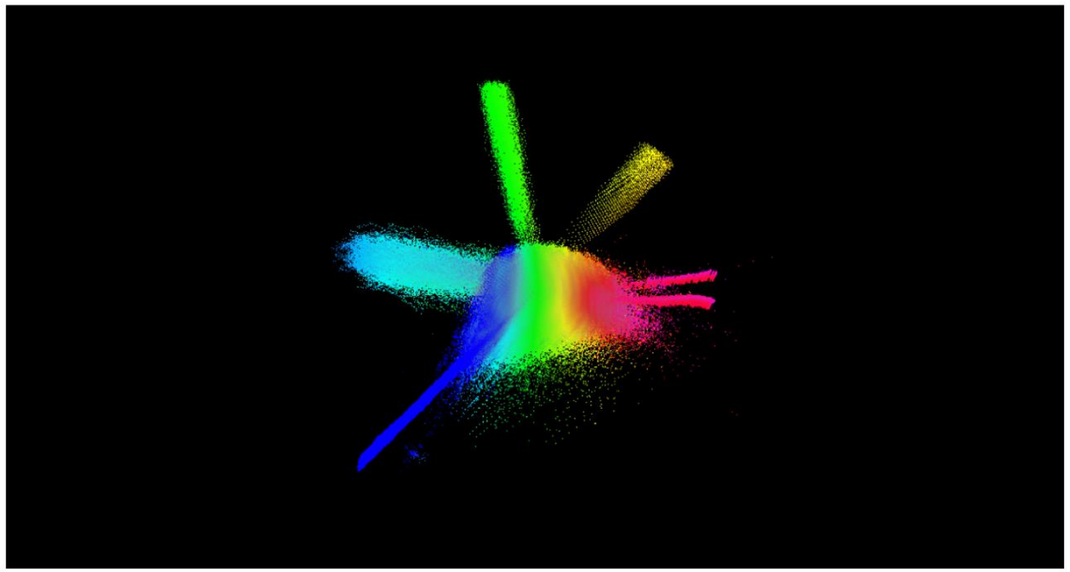
Clinical Trials

Predict clinical trials success rate, recognize the weak points in trial design, while adopting the best practices in the industry

NVIDIA Expands Large Language Models to Biology

Leading pharma companies, biotech startups and pioneering biology researchers are developing AI applications with the NVIDIA BioNeMo LLM service and framework to generate, predict and understand biomolecular data.

September 20, 2022 by ABRAHAM STERN



https://insilico.com/

NVIDIA

MATERIAL SCIENCE

Already changing related disciplines

DISCOVERY OF 2D MATERIALS USING TRANSFORMER NETWORK BASED GENERATIVE DESIGN *

Rongzhi Dong
Department of Computer Science and Engineering
University of South Carolina
Columbia, SC 29201

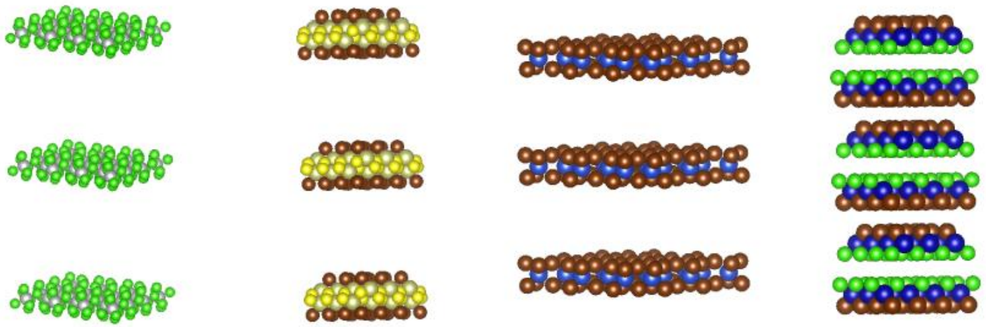
Yuqi Song
Department of Computer Science and Engineering
University of South Carolina
Columbia, SC 29201

Edirisuriya M. D. Siriwardane
Department of Physics
University of Colombo
Colombo 00300, Sri Lanka

Jianjun Hu *
Department of Computer Science and Engineering
University of South Carolina
Columbia, SC 29201
jianjunh@cse.sc.edu

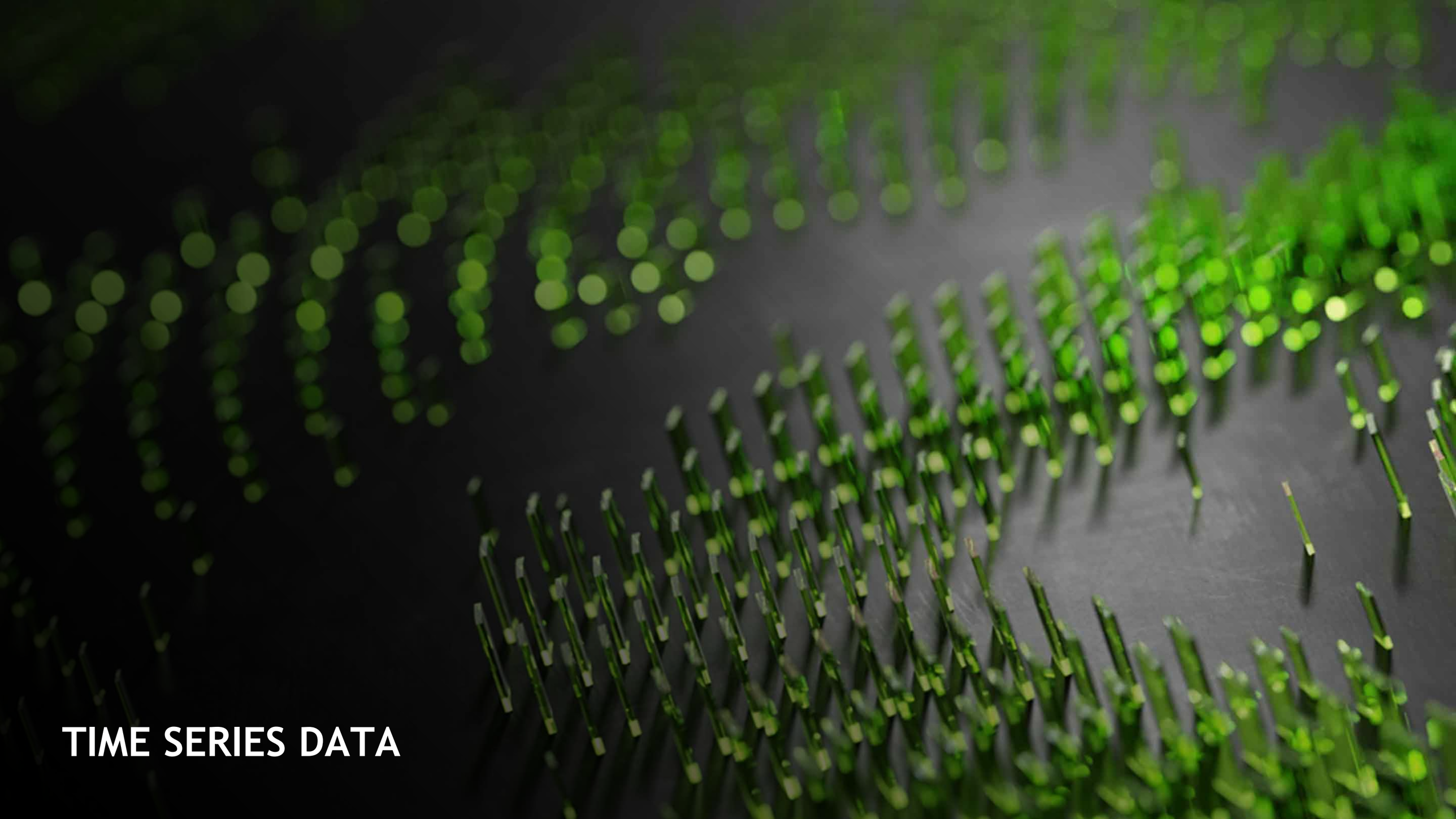
ABSTRACT

Two-dimensional (2D) materials have wide applications in superconductors, quantum, and topological materials. However, their rational design is not well established, and currently less than 6,000 experimentally synthesized 2D materials have been reported. Recently, deep learning, data-mining, and density functional theory (DFT)-based high-throughput calculations are widely performed to discover potential new materials for diverse applications. Here we propose a generative material design pipeline, namely material transformer generator(MTG), for large-scale discovery of hypothetical 2D materials. We train two 2D materials composition generators using self-learning neural language models based on Transformers with and without transfer learning. The models are then used to generate a large number of candidate 2D compositions, which are fed to known 2D materials templates for crystal structure prediction. Next, we performed DFT computations to study their thermodynamic stability based on energy-above-hull and formation energy. We report four new DFT-verified stable 2D materials with zero e-above-hull energies, including $NiCl_4$, IrSBr, $CuBr_3$, and CoBrCl. Our work thus demonstrates the potential of our MTG generative materials design pipeline in the discovery of novel 2D materials and other functional materials.



(a) $NiCl_4$ (b) IrSBr (c) $CuBr_3$ (d) CoBrCl

Figure 9: Four new 2D structures discovered by our MTG pipeline with 0 E-above-hull energy.


TIME SERIES DATA

BEYOND SPEECH

Foundation for a range of timeseries problems

”Voicebox is a non-autoregressive flow-matching model trained to infill speech, given audio context and text, trained on over 50K hours of speech that are neither filtered nor enhanced.”

BEYOND SPEECH

Taking the learnings to other disciplines

Predicting brain activity using Transformers

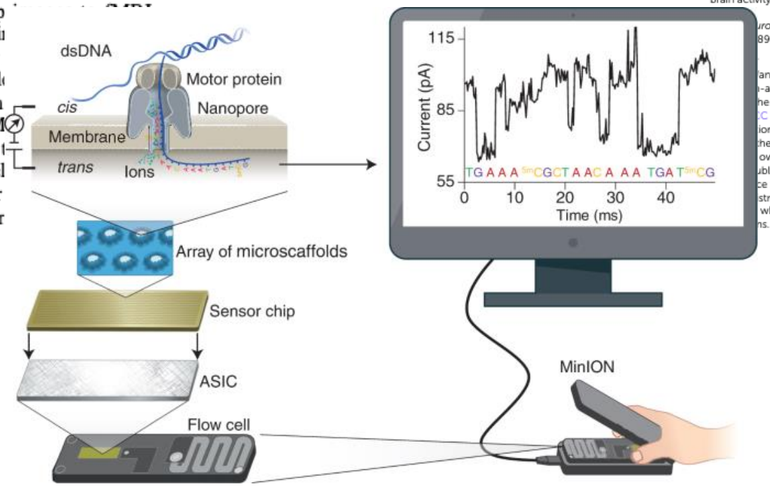
Hossein Adeli[1]*, **Sun Minni**[1], **Nikolaus Kriegeskorte**[1]

[1]Zuckerman Mind Brain Behavior Institute, Columbia University, New York, USA

* corresponding author: ha2366@columbia.edu

Abstract

The Algonauts challenge [Gifford et al., 2023] called on the community to provide novel solutions for predicting brain activity of humans viewing natural scenes. This report provides an overview and technical details of our submitted solution. We use a general transformer encoder-decoder model to map [illegible] responses. The encoder model is a vision transformer trained usi[illegible] methods (DINOv2). The decoder uses queries corresponding [illegible] regions of interests (ROI) in different hemispheres to gather rel[illegible] from the encoder output for predicting neural activity in each [illegible] tokens from the decoder are then linearly mapped to the fM[illegible] predictive success (challenge score: 63.5229, rank 2) suggests t[illegible] self-supervised transformers may deserve consideration as model[illegible] brain representations and shows the effectiveness of transformer [illegible] and cross-attention) to learn the mapping from features to brain r[illegible] available in this github repository.



frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 24 March 2023
DOI 10.3389/fnins.2023.1148855

Check for updates

OPEN ACCESS

EDITED BY
Xin Huang,
Renmin Hospital of Wuhan University, China

REVIEWED BY
Hongzhi Kuai,
Maebashi Institute of Technology, Japan
Yaofei Xie,
Xuzhou Medical University, China

*CORRESPONDENCE
Wenfeng Duan
ndyfy02345@ncu.edu.cn

SPECIALTY SECTION
This article was submitted to Visual Neuroscience, a section of the journal Frontiers in Neuroscience

RECEIVED 20 January 2023
ACCEPTED 06 March 2023
PUBLISHED 24 March 2023

CITATION
Wan Z, Li M, Liu S, Huang J, Tan H and Duan W (2023) EEGformer: A transformer–based brain activity classification method using EEG [illegible]rosci. 17:1148855. [illegible]89/fnins.2023.1148855

EEGformer: A transformer–based brain activity classification method using EEG signal

Zhijiang Wan[1,2,3], Manyu Li[2], Shichang Liu[4], Jiajin Huang[5], Hai Tan[6] and Wenfeng Duan[1]*

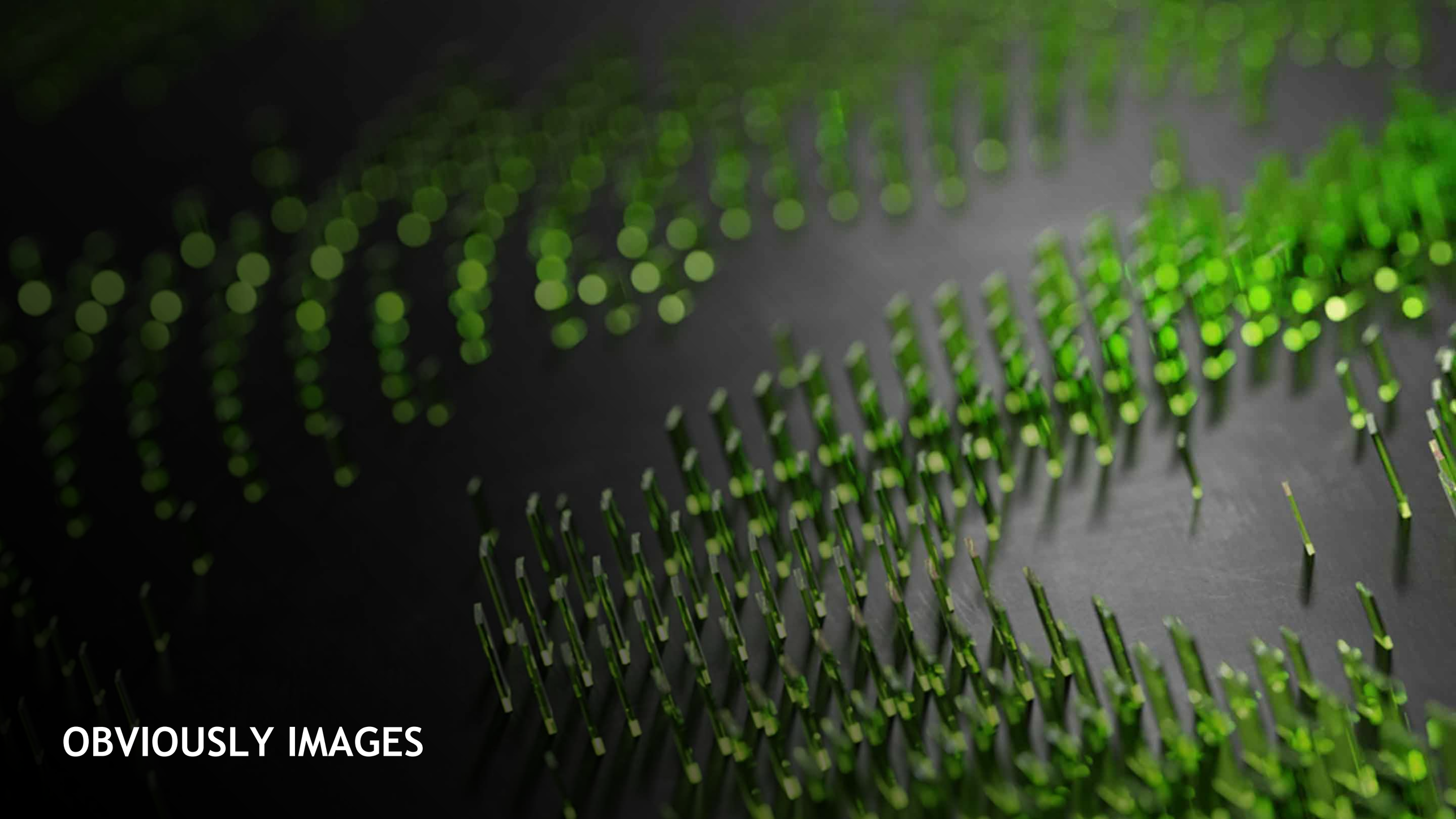
[1]The First Affiliated Hospital of Nanchang University, Nanchang University, Nanchang, Jiangxi, China, [2]School of Information Engineering, Nanchang University, Nanchang, Jiangxi, China, [3]Industrial Institute of Artificial Intelligence, Nanchang University, Nanchang, Jiangxi, China, [4]School of Computer Science, Shaanxi Normal University, Xi'an, Shaanxi, China, [5]Faculty of Information Technology, Beijing University of Technology, Beijing, China, [6]School of Computer Science, Nanjing Audit University, Nanjing, Jiangsu, China

Background: The effective analysis methods for steady-state visual evoked potential (SSVEP) signals are critical in supporting an early diagnosis of glaucoma. Most efforts focused on adopting existing techniques to the SSVEPs-based brain–computer interface (BCI) task rather than proposing new ones specifically suited to the domain.

Method: Given that electroencephalogram (EEG) signals possess temporal, regional, and synchronous characteristics of brain activity, we proposed a transformer–based EEG analysis model known as EEGformer to capture the EEG characteristics in a unified manner. We adopted a one-dimensional convolution neural network (1DCNN) to automatically extract EEG-channel-wise features. The output was fed into the EEGformer, which is sequentially constructed using three components: regional, synchronous, and temporal transformers. In addition to using a large benchmark database (BETA) toward SSVEP-BCI application to validate model performance, we compared the EEGformer to current state-of-the-art deep learning models using two EEG datasets, which are obtained from our previous study: SJTU emotion EEG dataset (SEED) and a depressive EEG database (DepEEG).

Results: The experimental results show that the EEGformer achieves the best classification performance across the three EEG datasets, indicating that the rationality of our model architecture and learning EEG characteristics in a unified manner can improve model classification performance.

Conclusion: EEGformer generalizes well to different EEG datasets, demonstrating our approach can be potentially suitable for providing accurate brain activity classification and being used in different application scenarios, such as SSVEP-based early glaucoma diagnosis, emotion recognition and depression discrimination


OBVIOUSLY IMAGES

GENERATIVE MODELS

We understood how to design those for quite some time

THE NUCLEUS

Period of early success lays the foundation for the future of generative models.

GAN EXPLOSION

Success of Generative Adversarial Networks pushes the boundary of what is possible.

STABILITY AND SCALE

Working towards stable training of larger and more capable models.

FIDELITY

Successs in generation of higher fidelity content

REALISM

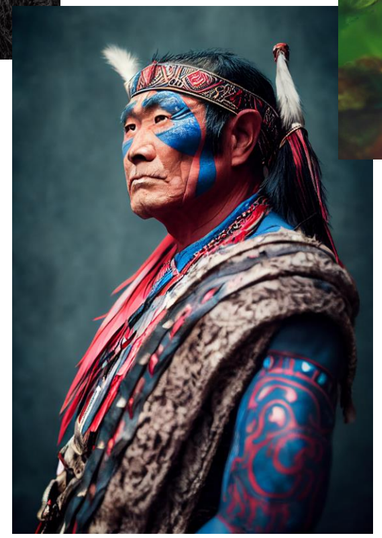
Incremental improvements increasing the realism of the generated content.

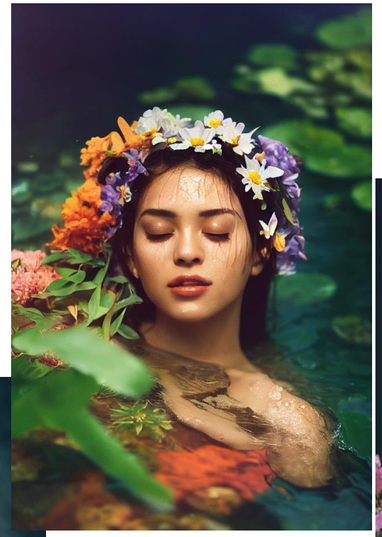
DIVERSITY AND CONTROL

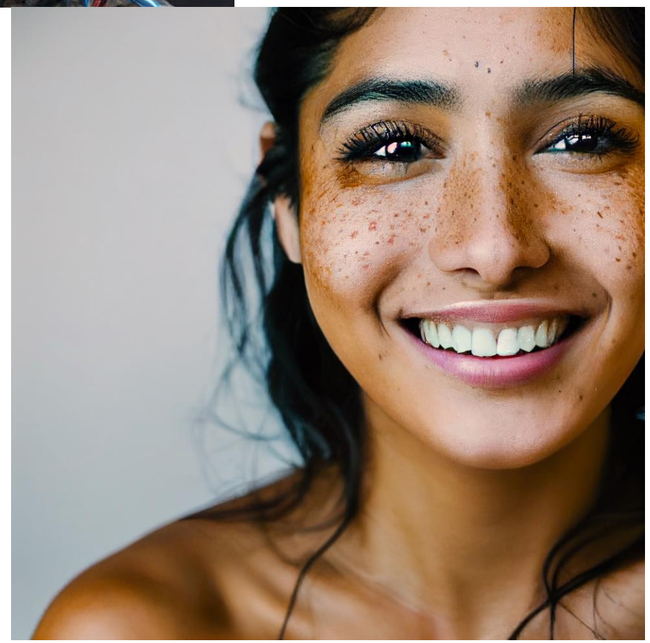
Models that not only generate high fidelity but also diverse content that can be controlled by the user.

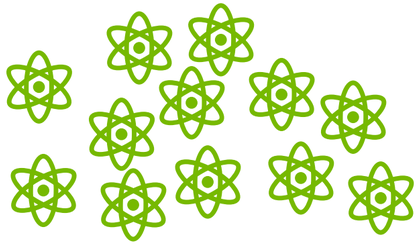
EVEN MORE DIVERSITY AND CONTROL

Blurring the line between digitally created art and reality





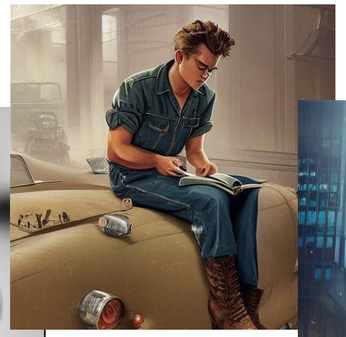
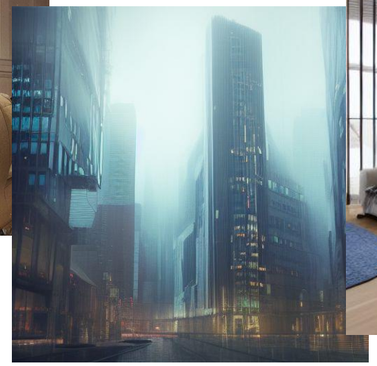
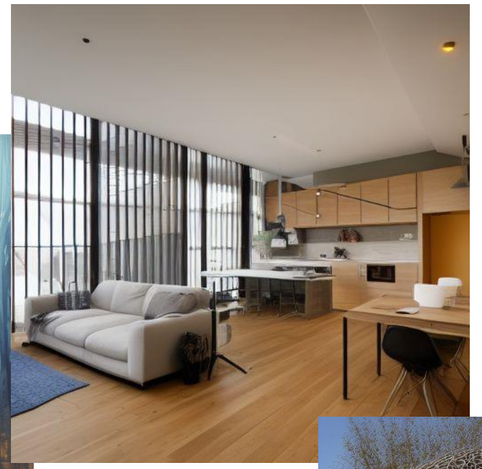
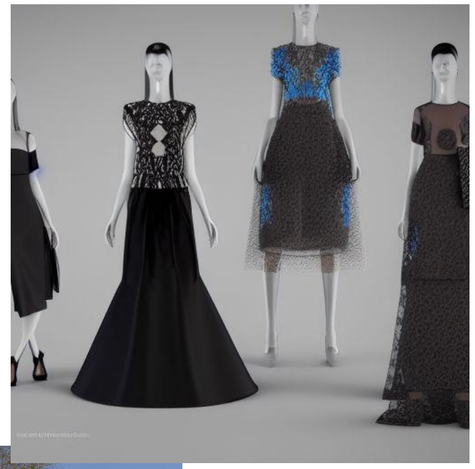
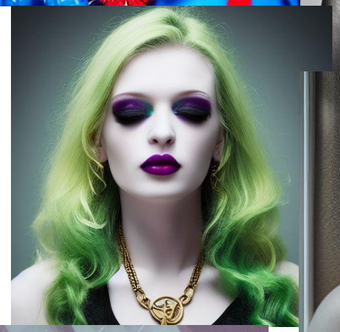
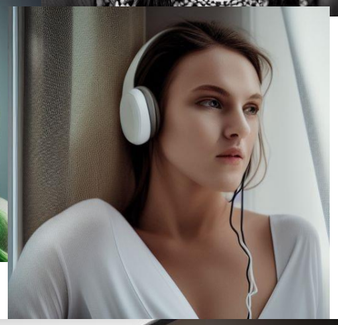
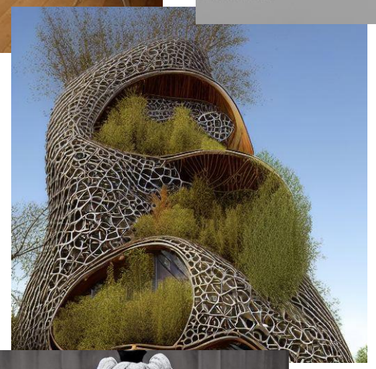


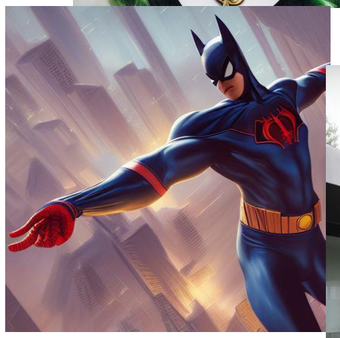
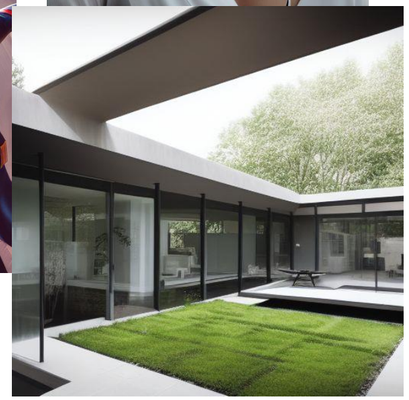


EASE OF USE

Critical mass



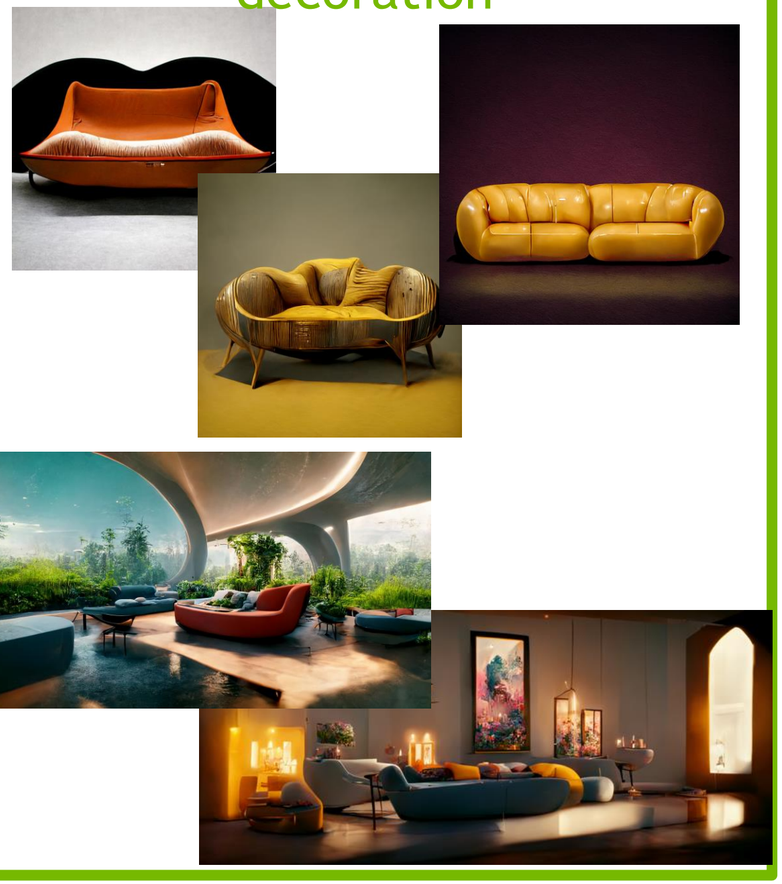








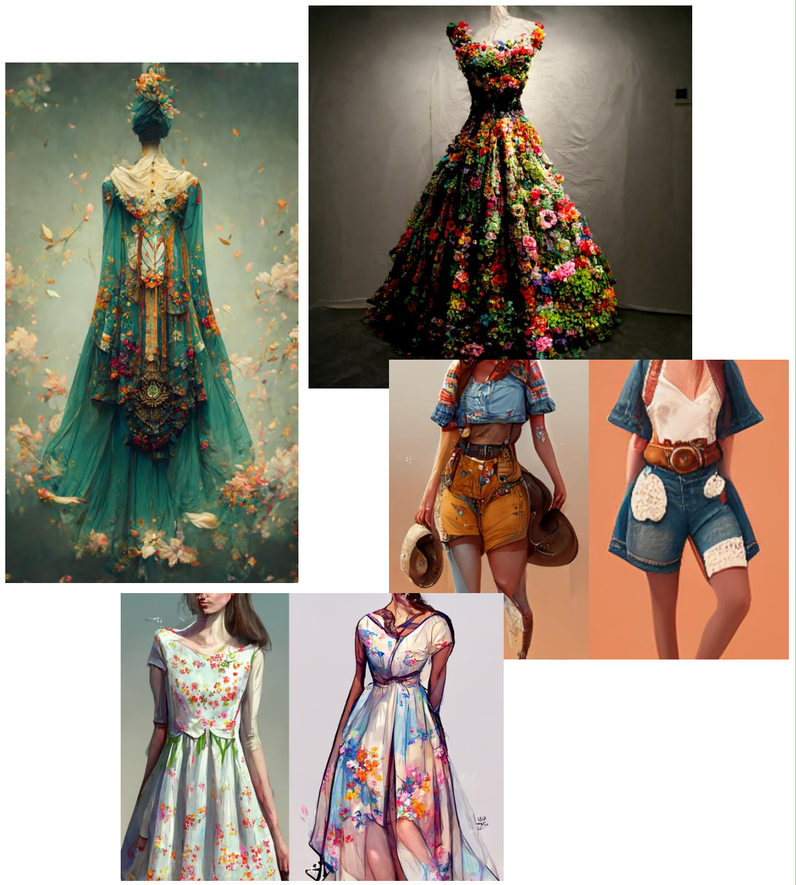

ANY FORM OF DESIGN

From Interior decoration to… Architecture

Furniture and interior decoration



Fashion



Architecture

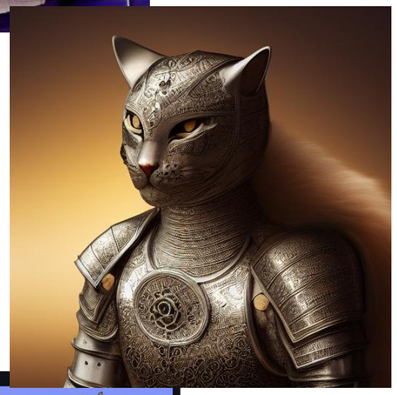
New York Times: A.I.-Generated Art Is Already Transforming Creative Work
https://www.nytimes.com/2022/10/21/technology/ai-generated-art-jobs-dall-e-2.html

ANY FORM OF DESIGN

…to Automotive and more

Automotive

Game development



Text to image on Stable Diffusion, using the prompt: 'magical off world dreamscape'

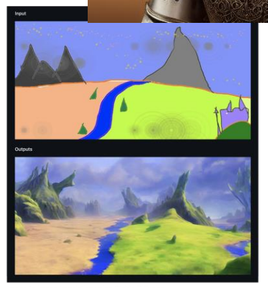
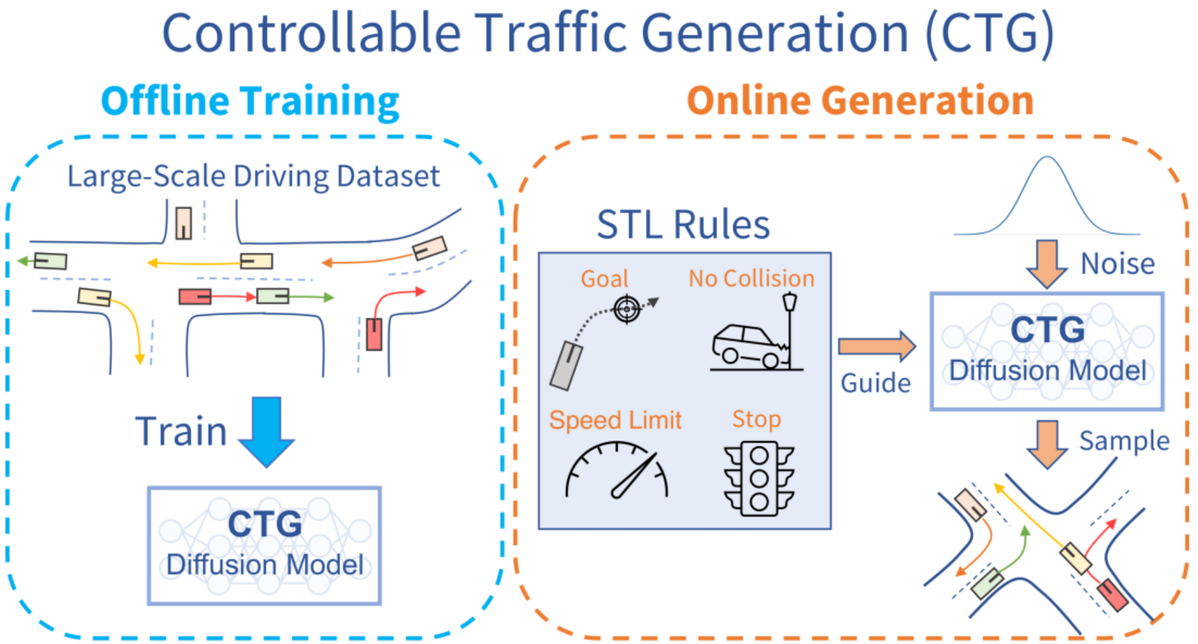


Image-to-image translation using Stable Diffusion

Biology / Chemistry / Material Science / Scientific Visualization / ???

SIMULATION

Guided Conditional Diffusion for Controllable Traffic Simulation



Controllable and realistic traffic simulation is critical for developing and verifying autonomous vehicles. Typical heuristic-based traffic models offer flexible control to make vehicles follow specific trajectories and traffic rules. On the other hand, data-driven approaches generate realistic and human-like behaviors, improving transfer from simulated to real-world traffic. However, to the best of our knowledge, no traffic model offers both controllability and realism. In this work, we develop a conditional diffusion model for controllable traffic generation (CTG) that allows users to control desired properties of trajectories at test time (e.g., reach a goal or follow a speed limit) while maintaining realism and physical feasibility through enforced dynamics. The key technical idea is to leverage recent advances from diffusion modeling and differentiable logic to guide generated trajectories to meet rules defined using signal temporal logic (STL). We further extend guidance to multi-agent settings and enable interaction-based rules like collision avoidance. CTG is extensively evaluated on the nuScenes dataset for diverse and composite rules, demonstrating improvement over strong baselines in terms of the controllability-realism tradeoff.

TRANSFORMING DATA PROCESSING

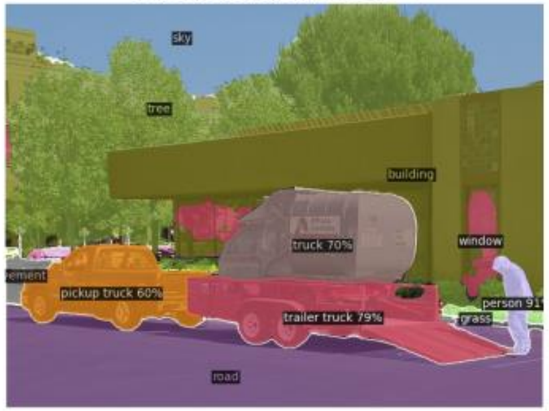
Again dramatically reducing the skills barrier

Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models

Input Image

K-Means Clustering of Internal Diffusion Features

Open-Vocabulary Panoptic Segmentation Prediction from ODISE



We present ODISE: Open-vocabulary DIffusion-based panoptic SEgmentation, which unifies pre-trained text-image diffusion and discriminative models to perform open-vocabulary panoptic segmentation. Text-to-image diffusion models have the remarkable ability to generate high-quality images with diverse open-vocabulary language descriptions. This demonstrates that their internal representation space is highly correlated with open concepts in the real world. Text-image discriminative models like CLIP, on the other hand, are good at classifying images into open-vocabulary labels. We leverage the frozen internal representations of both these models to perform panoptic segmentation of any category in the wild. Our approach outperforms the previous state of the art by significant margins on both open-vocabulary panoptic and semantic segmentation tasks. In particular, with COCO training only, our method achieves 23.4 PQ and 30.0 mIoU on the ADE20K dataset, with 8.3 PQ and 7.9 mIoU absolute improvement over the previous state of the art. We open-source our code and models at https://github.com/NVlabs/ODISE.

TRANSFORMING DATA COLLECTION

Automotive example

ROBOTICS

Planning and Imagination

Publications › StructDiffusion: Language-Guided Creation of Physically-Valid Structures using Unseen Objects

StructDiffusion: Language-Guided Creation of Physically-Valid Structures using Unseen Objects

Robots operating in human environments must be able to rearrange objects into semantically-meaningful configurations, even if these objects are previously unseen. We focus on the problem of building physically-valid structures without step-by-step instructions.

We propose StructDiffusion, which combines a diffusion model and an object-centric transformer to construct structures given partial-view point clouds and high-level language goals, such as *"set the table"* and *"make a line"*.

StructDiffusion improves success rate on assembling physically-valid structures out of unseen objects by on average 16% over an existing multi-modal transformer model, while allowin us to use one multi-task model to produce a wider range of different structures. We show experiments on held-out objects in both simulation and on real-world rearrangement tasks.

PROGPROMPT: Generating Situated Robot Task Plans using Large Language Models

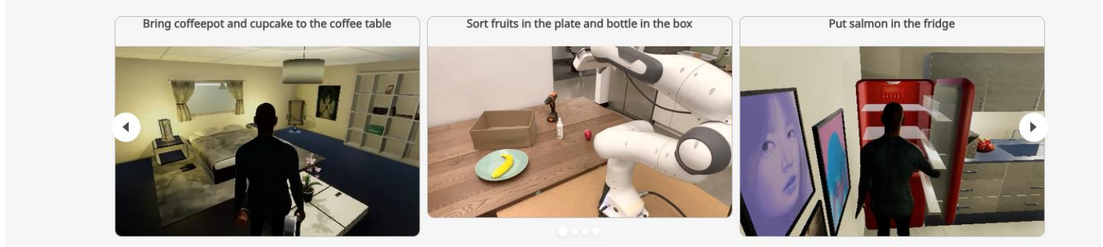
ICRA 2023

Extended version in Autonomous Robots 2023

Ishika Singh[1], Valts Blukis[2], Arsalan Mousavian[2], Ankit Goyal[2], Danfei Xu[2], Jonathan Tremblay[2], Dieter Fox[2], Jesse Thomason[1], Animesh Garg[2]

[1]University of Southern California, [2]NVIDIA

arXiv | Video | Code | Poster | FAQs



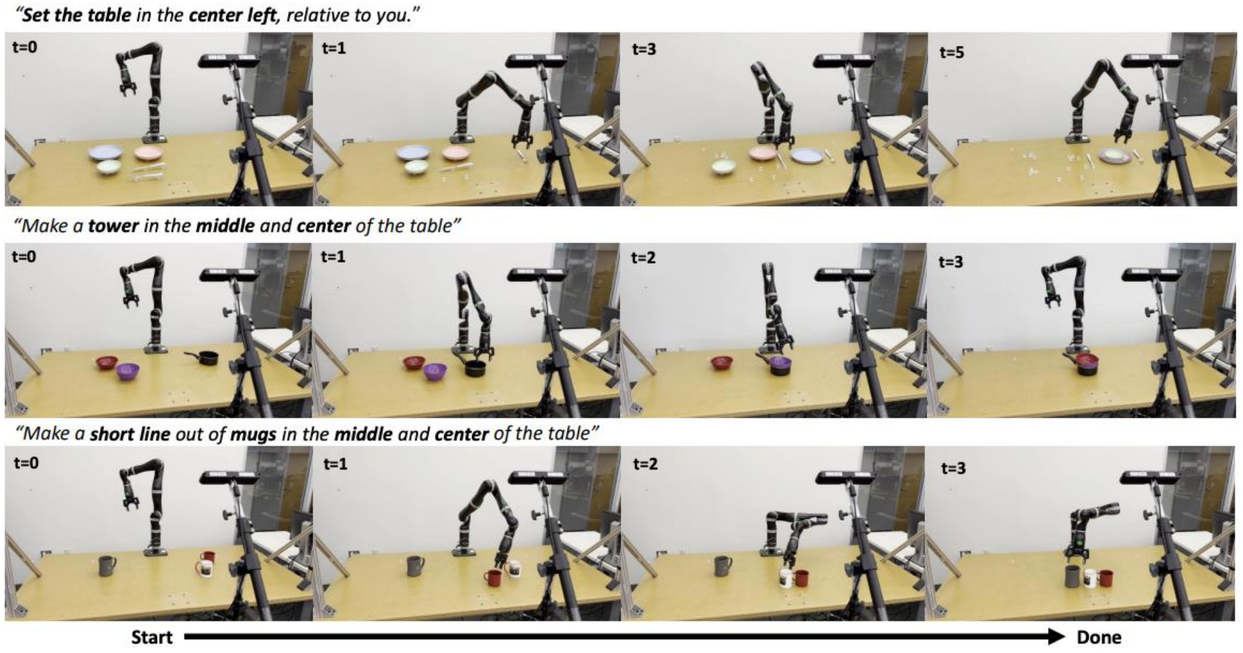


Fig. 1: Real-world rearrangement with unseen objects, given a language instruction. We use StructDiffusion to predict possible goals that satisfy physical constraints such as avoiding collisions between objects. At the core of StructDiffusion is an object-centric multimodal transformer backbone combined with a diffusion model, capable of sampling diverse high-level motion goals for language-guided rearrangement.

NVIDIA

PHYSICS

A Physics-informed Diffusion Model for High-fidelity Flow Field Reconstruction

Dule Shu,[†,§] Zijie Li,[†,§] and Amir Barati Farimani[*,†,‡,¶]

†*Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh PA, USA*

‡*Machine Learning Department, Carnegie Mellon University, Pittsburgh PA, USA*

¶*Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh PA, USA*

§*Contributed equally to this work*

E-mail: barati@cmu.edu

Abstract

Machine learning models are gaining increasing popularity in the domain of fluid dynamics for their potential to accelerate the production of high-fidelity computational fluid dynamics data. However, many recently proposed machine learning models for high-fidelity data reconstruction require low-fidelity data for model training. Such requirement restrains the application performance of these models, since their data reconstruction accuracy would drop significantly if the low-fidelity input data used in model test has a large deviation from the training data. To overcome this restraint, we propose a diffusion model which only uses high-fidelity data at training. With different configurations, our model is able to reconstruct high-fidelity data from either a regular low-fidelity sample or a sparsely measured sample, and is also able to gain an accuracy increase by using physics-informed conditioning information from a known partial differential equation when that is available. Experimental results demonstrate that our model can produce accurate reconstruction results for 2d turbulent flows based on different input sources without retraining.

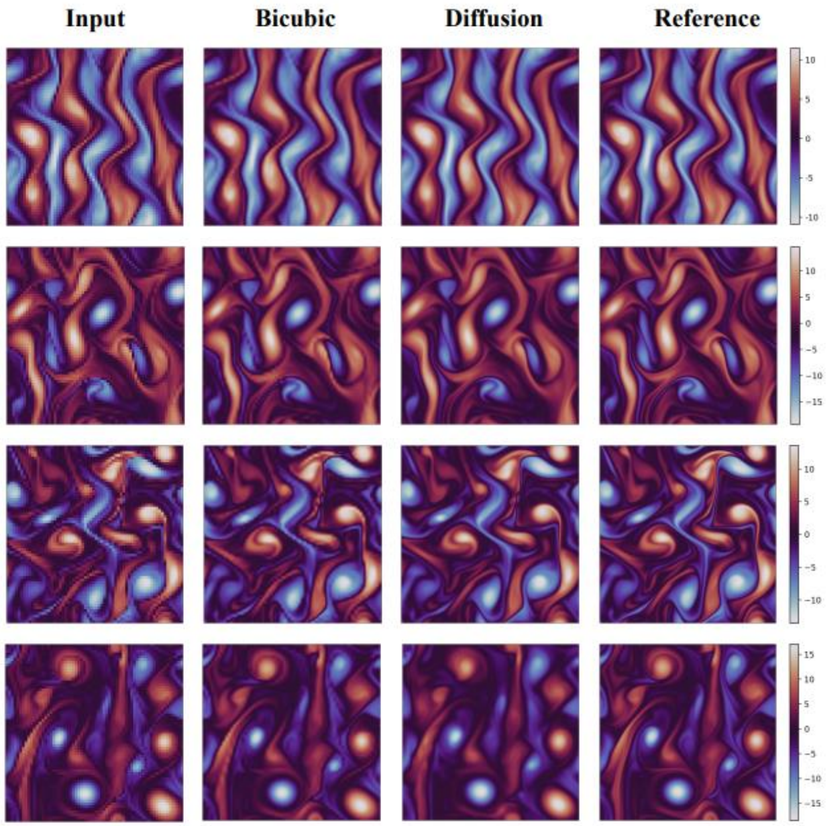
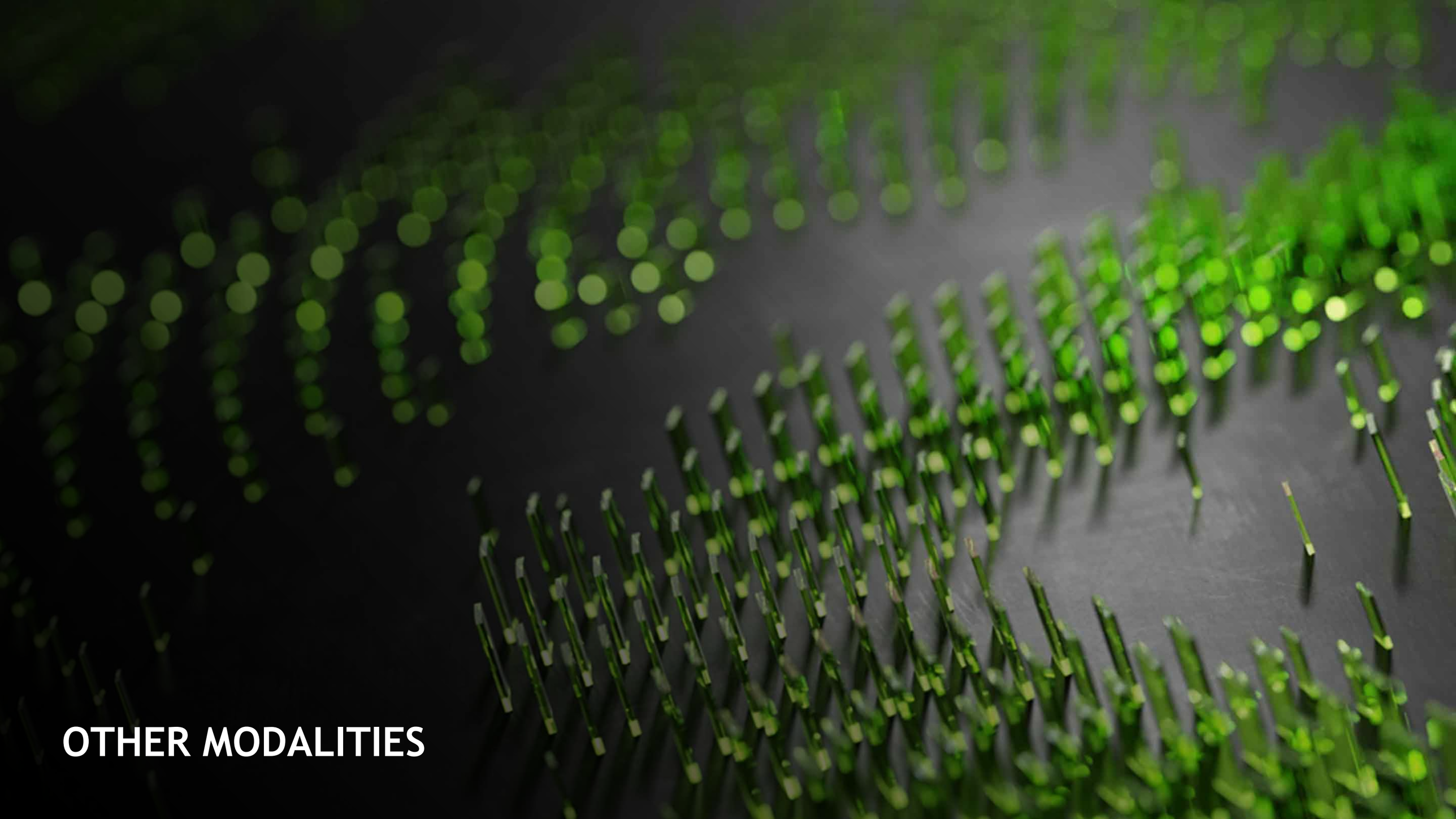


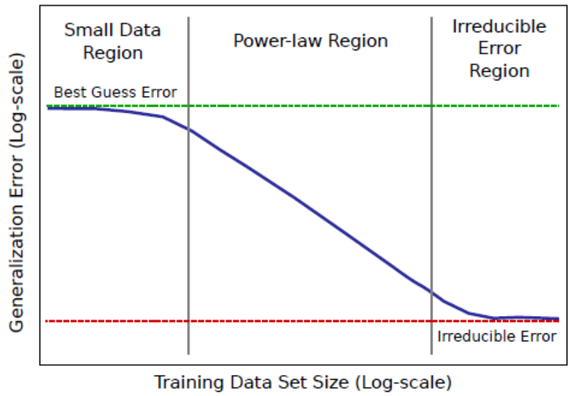
Figure 3: Qualitative comparison of different upsampling methods on 4x upsampling task.

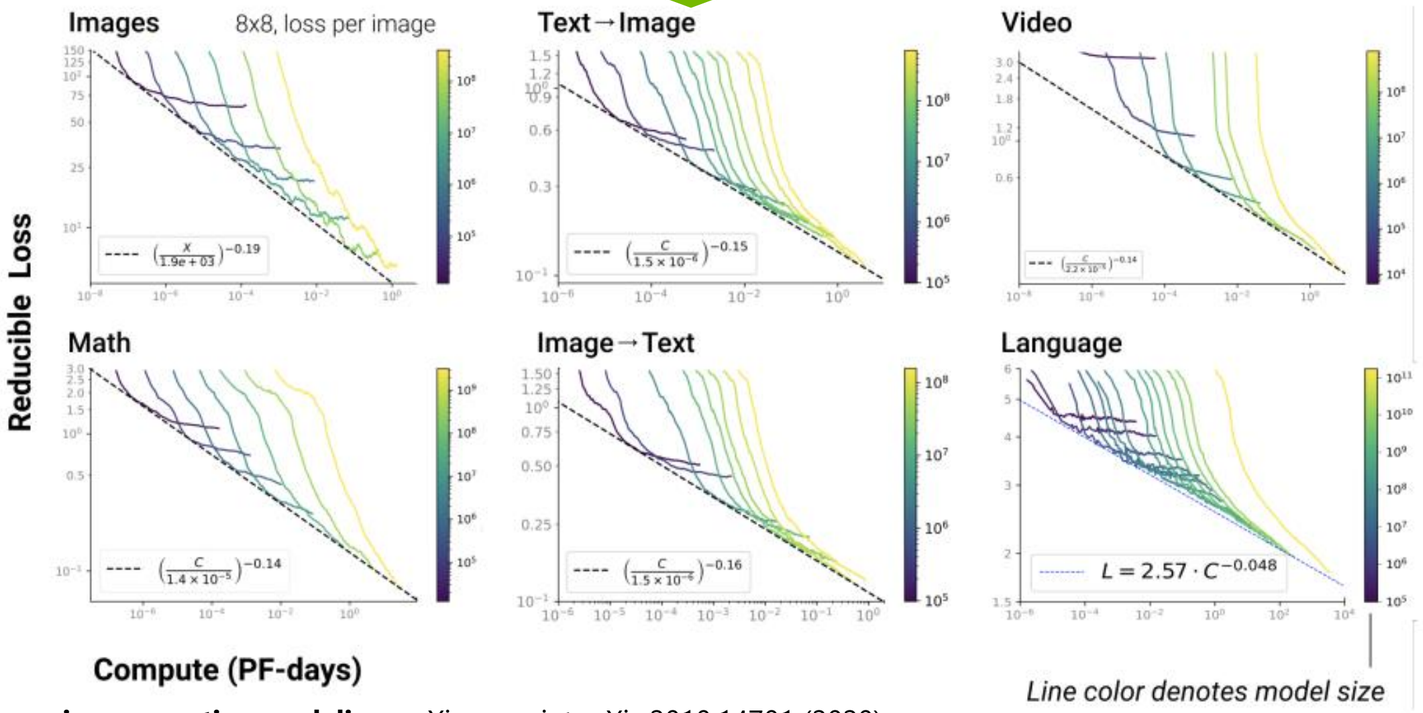


OTHER MODALITIES

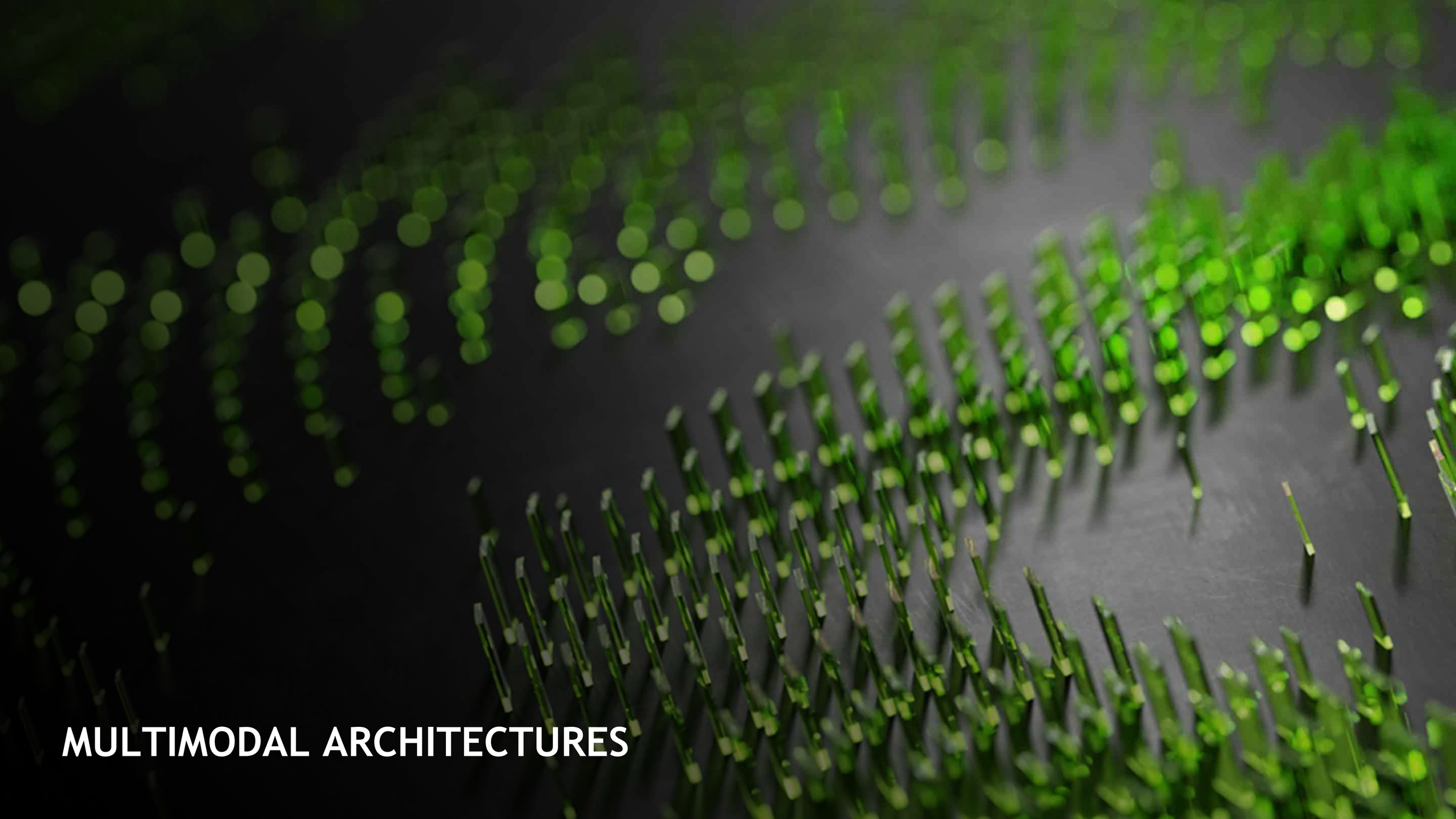
EMPIRICAL EVIDENCE

The Scaling Laws for Generative models



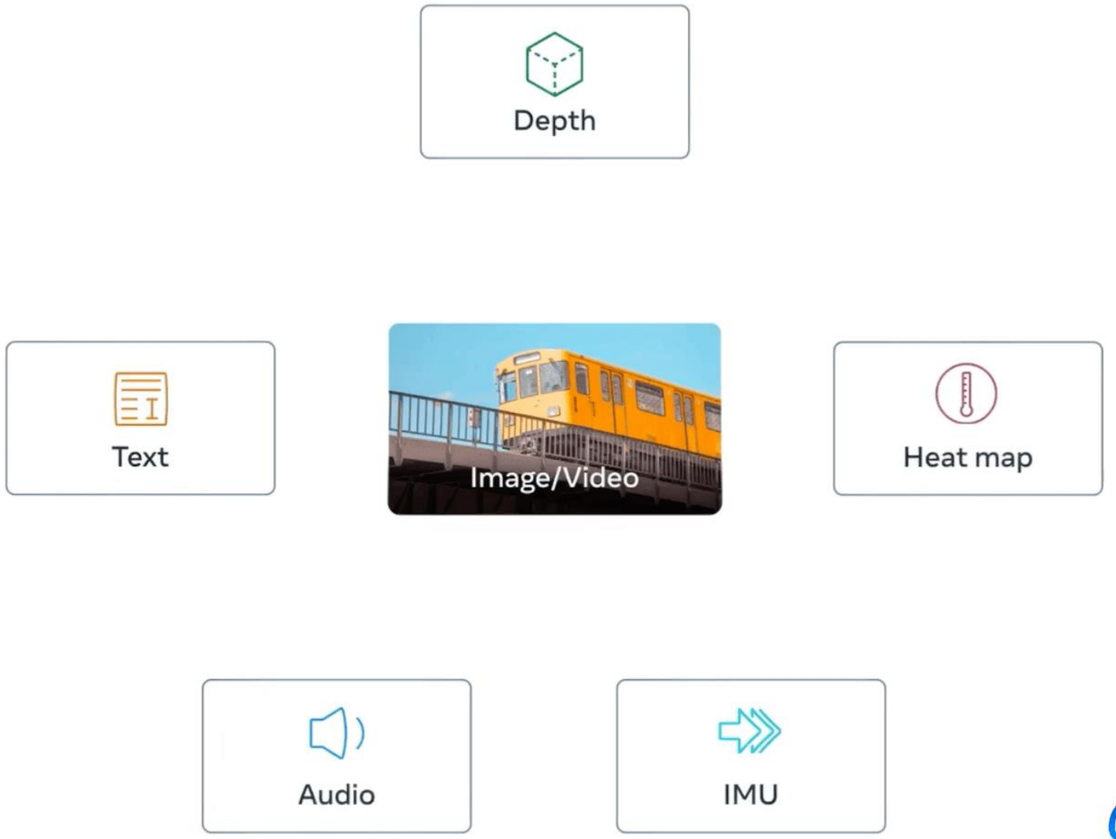


Henighan, Tom, et al. **Scaling laws for autoregressive generative modeling.** arXiv preprint arXiv:2010.14701 (2020).


MULTIMODAL ARCHITECTURES

This is just the first wave

Rise of multimodal architectures



Meta AI

Simplicity of multimodal architectures

LLAVA example

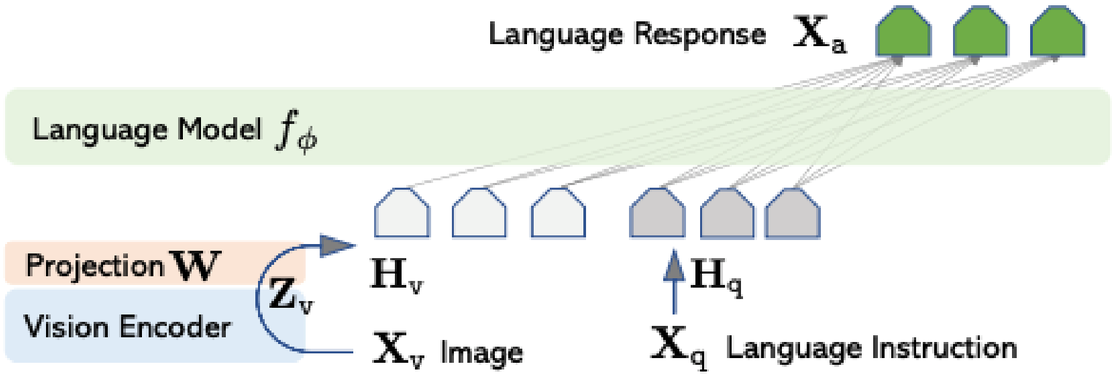
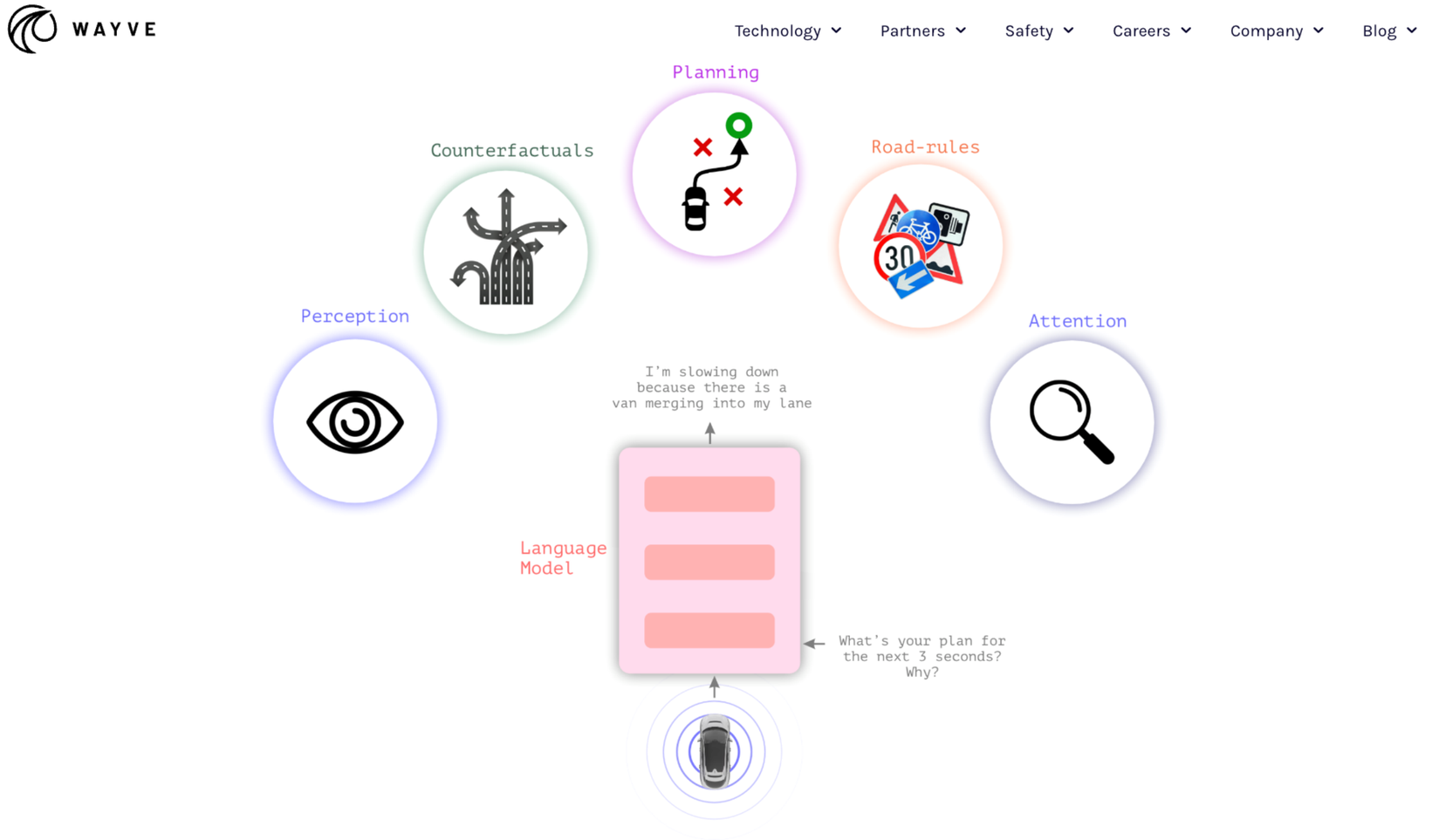
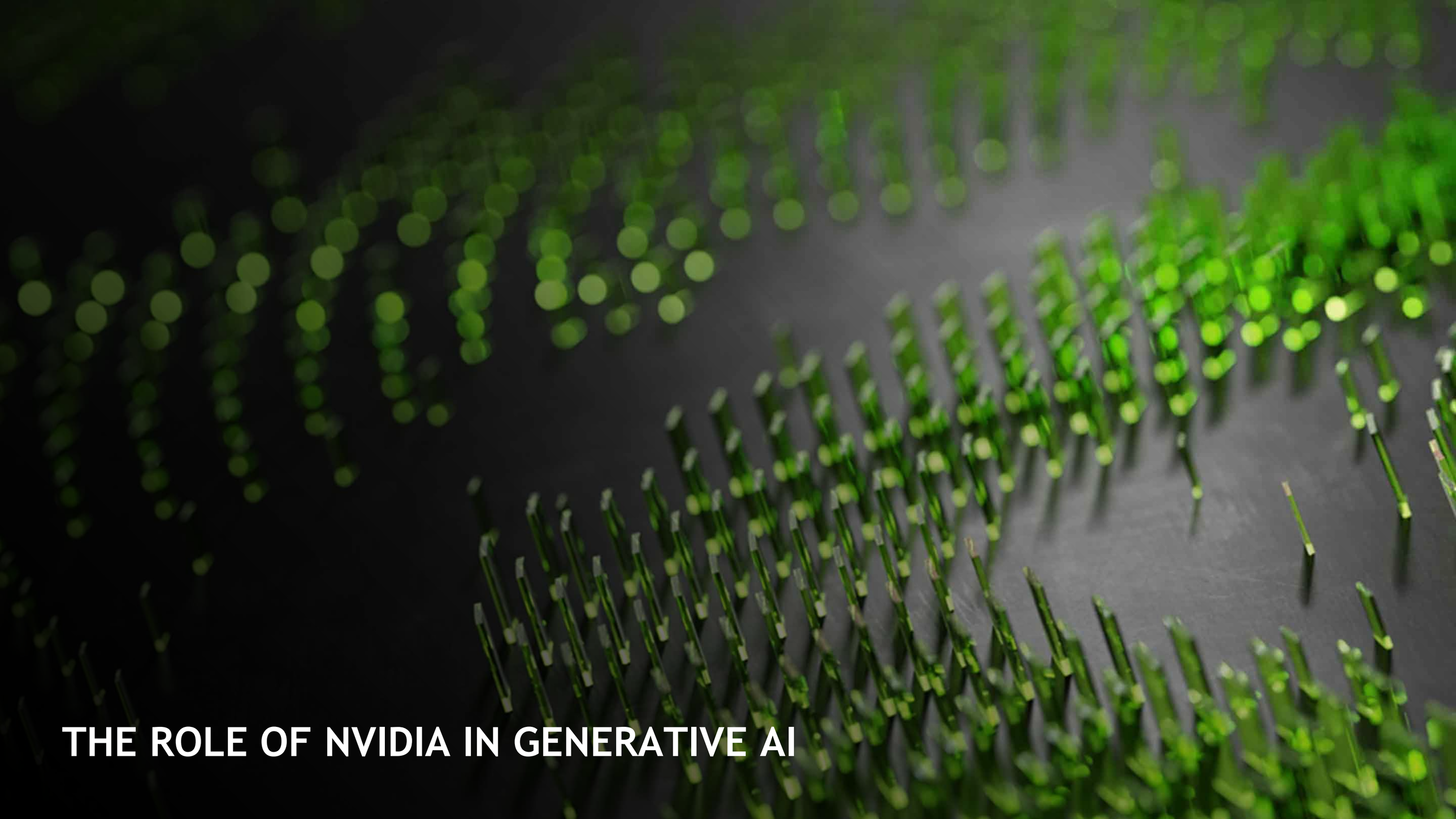


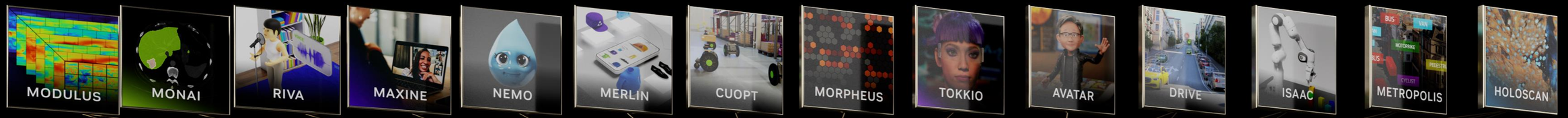
Figure 1: LLaVA network architecture.

This is just the first wave

Rise of multimodal architectures




THE ROLE OF NVIDIA IN GENERATIVE AI



AI APPLICATION FRAMEWORK

PLATFORMS

NVIDIA HPC

NVIDIA AI

NVIDIA Omniverse

ACCELERATION LIBRARIES

cuNumeric | CV-CUDA | cuQuantum | Parabricks | Sionna | JetPack

RAPIDS | Spark | cuDNN | cuGraph | TensorRT | Triton | DeepStream | Flare

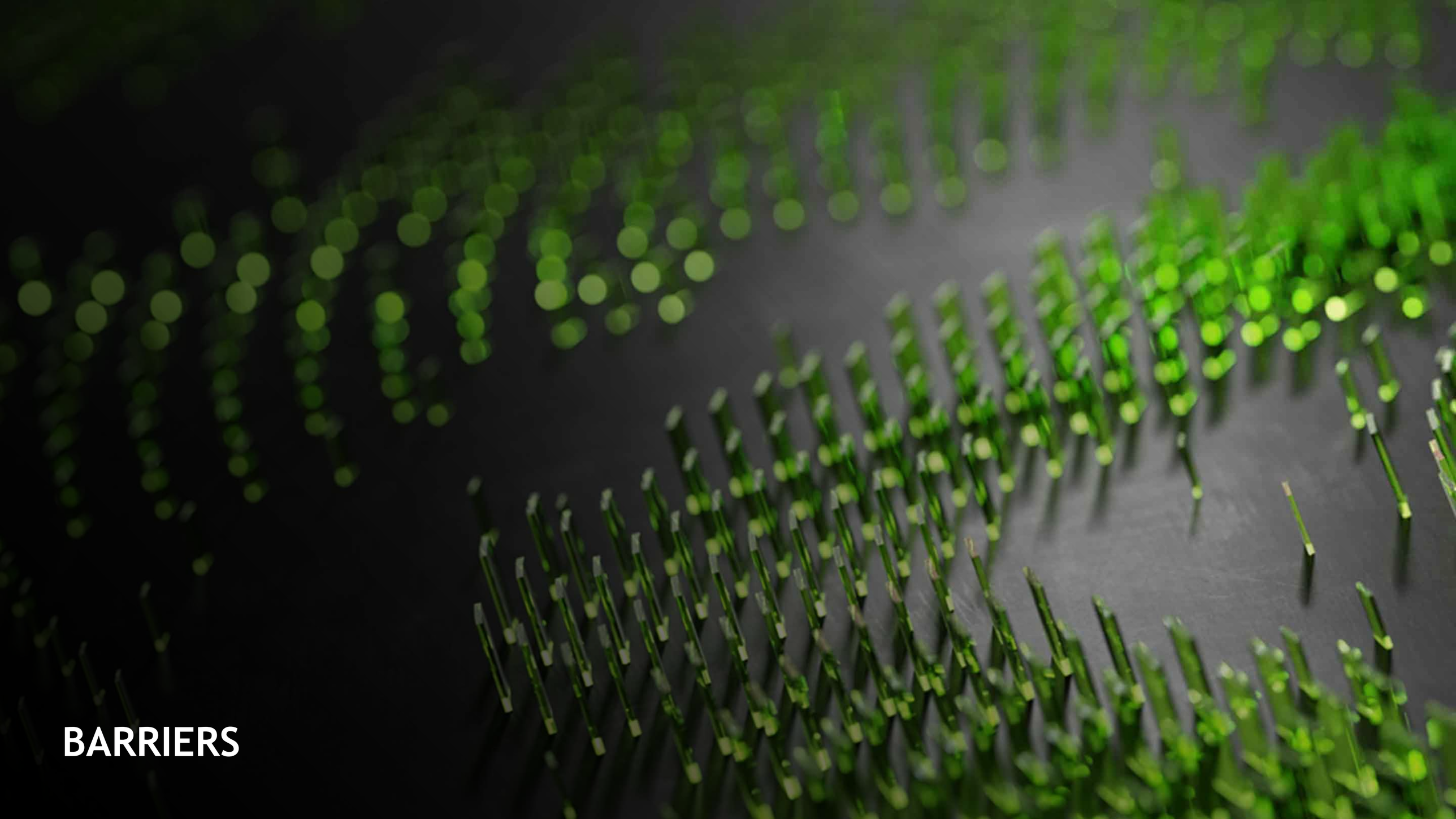
DOCA | Mag IO | Aerial

CLOUD-TO-EDGE DATACENTER-TO-ROBOTIC SYSTEMS

RTX | DGX | HGX | EGX | OVX | Super POD | AGX

3 CHIPS

GPU | CPU | DPU



BARRIERS

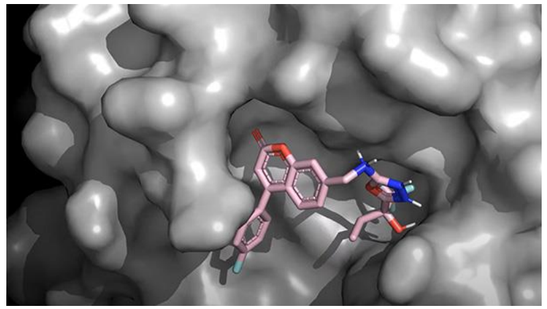

IS IT COMPLEXITY?

NVIDIA'S GENERATIVE AI SOLUTIONS

Foundations to Build and Run Your Generative AI

NVIDIA NeMo service

NVIDIA BioNeMo service

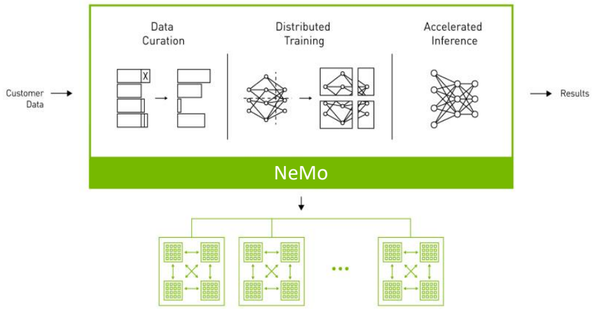


NVIDIA Picasso service

A photo of a golden retriever puppy wearing a green shirt. The shirt has text that says 'NVIDIA rocks". Background office. 4k dslr.

NVIDIA AI Foundations

NVIDIA NeMo framework

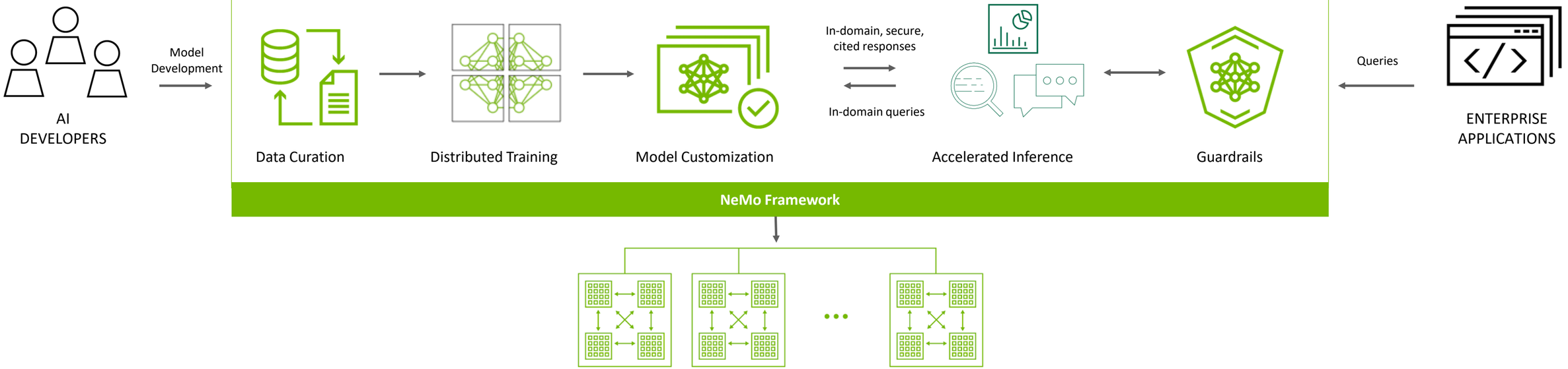


NVIDIA AI Enterprise

NVIDIA DGX Cloud

NEMO FRAMEWORK

End-to-end, cloud-native framework to build, customize and deploy generative AI models



Multi-modality support

Build language, image, generative AI models

Data Curation @ Scale

Extract, deduplicate, filter info from large unstructured data @ scale

Optimized Training

Accelerate training and throughput by parallelizing the model and the training data across 1,000s of nodes.

Model Customization

Easily customize with P-tuning, SFT, Adapters, RLHF, AliBi

Deploy at-scale Anywhere

Run optimized inference at-scale anywhere

Guardrails

Keep applications aligned with safety and security requirements using NeMo Guardrails

Support

NVIDIA AI Enterprise and experts by your side to keep projects on track

Now in open beta, general availability with NVIDIA AI Enterprise in Q2’2023 (LLMs Only)

Multi-modal available via early access now

DEMONSTRATING SIMPLICITY OF USE

Step 1

Formatting the dataset

Step 1

Alternative examples


```
{"taskname": "sentiment", "sentence": "Super lekarz i cz\u0142owiek przez du\u017ce C . Bardzo du\u017ce do\u015bwiadczenie i trafne diagnozy . Wielka cierpliwo\u015b\u0107 do ludzi starszych . Od lat opiekuje si\u0119 moj\u0105 Mam\u0105 staruszk\u0105 , i twierdz\u0119 , \u017ce mamy du\u017ce szcz\u0119\u015bcie , \u017ce mamy takiego lekarza . Naprawd\u0119 nie wiem coby\u015bmy zrobili , gdyby nie Pan doktor . Dzi\u0119ki temu , moja mama \u017cyje . Ka\u017cda wizyta u specjalisty jest u niego konsultowana i uwa\u017cam , \u017ce jest lepszy od ka\u017cdego z nich . Mamy do Niego prawie nieograniczone zaufanie . Mo\u017cna wiele dobrego o Panu doktorze jeszcze napisa\u0107 . Niestety , ma bardzo du\u017co pacjent\u00f3w , jest przepracowany ( z tego powodu nawet obawiam si\u0119 o jego zdrowie ) i dost\u0119p do niego jest trudny , ale zawsze mo\u017cliwy .", "label": "Pozytywny"}
{"taskname": "sentiment", "sentence": "Bardzo olewcze podejscie do pacjenta . Przyprowadzajac dziecko z ostra wysypka na calym ciele trwajaca od 2 tygodni Pani doktor stwierdzila ze nie widzi wskazan wystawienia dziecku skierowania na testy skorne . Chocby na nasza prosbe i dodam iz w prywatnej klinice ( gdzie tak czy siak musielibysmy za to zaplacic ) Odmowila rowniez wystawienia zaswiadczenia o tym ze dziecko jest \" zdrowe \" i moze uczeszczac do przedszkola twierdzac ze takich zaswiadczen sie nie wystawia . L4 oczywiscie rozniez nie wchodzilo w gre , . Jednym slowem z gabinetu wyszlismy bez niczego mimo iz kierowani bylismy przez dwoch innych pediatrow na natychmiastowe testy do alergologa .", "label": "Negatywny"}
{"taskname": "sentiment", "sentence": "Lekarz zaleci\u0142 mi kuracj\u0119 alternatywn\u0105 do dotychczasowej , wi\u0119c jeszcze nie daj\u0119 najwy\u017cszej oceny ( zobaczymy na ile oka\u017ce si\u0119 skuteczna ) . Do Pana doktora nie mam zastrze\u017ce\u0144 : bardzo profesjonalny i kulturalny . Jedyny minus dotyczy gabinetu , kt\u00f3ry nie jest nowoczesny , co mo\u017ce zniech\u0119ca\u0107 pacjentki .", "label": "Nieznany"}
{"taskname": "sentiment", "sentence": " Konsumenci oczywi\u015bcie kieruj\u0105 si\u0119 cen\u0105 . Te leki s\u0105 ta\u0144sze , ale dzisiaj nie mo\u017cemy ju\u017c powiedzie\u0107 , \u017ce znacznie ta\u0144sze . Jest to mniej kr\u0119puj\u0105ce , nie trzeba wychodzi\u0107 z domu , nie trzeba udawa\u0107 si\u0119 do lekarza czy prosi\u0107 o recept\u0119 . W takich przypadkach kupuj\u0105cy nie kieruje si\u0119 zdrowym rozs\u0105dkiem , ale cen\u0105 – naiwnie szukaj\u0105c lek\u00f3w z niesprawdzonych \u017ar\u00f3de\u0142 . Podrobione leki mog\u0105 by\u0107 ska\u017cone , zanieczyszczone lub zawiera\u0107 substancje toksyczne czy w og\u00f3le nieodpowiedni sk\u0142ad . Cz\u0119sto maj\u0105 te\u017c sk\u0142adniki nieaktywne – w \u017caden spos\u00f3b nie wp\u0142ywaj\u0105ce na popraw\u0119 b\u0105d\u017a ( powoduj\u0105ce ) pogorszenie stanu zdrowia – wyja\u015bni\u0142a ekspertka . Doda\u0142a tak\u017ce , \u017ce \" podr\u00f3bki \" mog\u0105 mie\u0107 r\u00f3wnie\u017c sk\u0142adniki aktywne w nieodpowiednich dawkach , z byt silne czy zmieszane w spos\u00f3b nieodpowiedni , a badania pokazuj\u0105 , \u017ce w podrobionych lekach znajduj\u0105 si\u0119 trucizny dla szczur\u00f3w , wosk do pod\u0142ogi , cement , gips i podobne substancje .", "label": "Neutralny"}
{"taskname": "sentiment", "sentence": "Pani Doktor Iwona jest profesjonalistk\u0105 w ka\u017cdym calu : ) Id\u0105c do stomatologa zawsze czuj\u0119 przera\u017aliwy strach . Pani Doktor Iwona jest wyrozumia\u0142a i delikatna , bardzo mi\u0142a , cierpliwie wyja\u015bnia ka\u017cdy zaistnia\u0142y problem a co najbardziej mi si\u0119 podoba to fakt , \u017ce w swoim dzia\u0142aniu jest bardzo stanowcza i konkretna . Dla mnie to jasny znak , \u017ce specjalista kt\u00f3ry si\u0119 mn\u0105 zajmuje , doskonale wie co i jak robi . Pozwala mi to poczu\u0107 si\u0119 bezpiecznie . Wiem , \u017ce jestem w dobrych r\u0119kach . A do tego u\u015bmiech i ciep\u0142o . . . . . po strachu ani \u015bladu . Pani Iwono DZI\u0118KUJ\u0118 : D", "label": "Pozytywny"}
{"taskname": "sentiment", "sentence": "Jest nie prawda co napisal ten internauta . Ten lekarz jest bardzo dobrym , przesympatycznym , milym i kulturalnym czlowiekiem . Nie jest tez prawda ze oglada kobiece piersi , wcale tego nie robi , aczkolwiek w wielu przypadkach powinien to robic , moze uratowal by jakas kobiete od smierci . Lecze sie u Pana doktora juz 8 lat i nie moge powiedziec na niego zadnego zlego slowa , wrecz przeciwnie . zarowno On sam jak i personel tej przychodni to wspaniali ludzie . Goraco polecam .", "label": "Pozytywny"}
{"taskname": "sentiment", "sentence": "Krzysztof jest ZNAKOMITYM fizjoterapeut\u0105 ! Przysz\u0142a m do niego z wypadaj\u0105cym , krzywym kolanem , po operacji wi\u0119zad\u0142a oraz z bol\u0105c\u0105 \u0142\u0119kotk\u0105 w drugim kolanie . Krzysztof po wywiadzie og\u00f3lnym oraz badaniu stwierdzi\u0142 szybko co mo\u017ce na to pom\u00f3c i jak temu przeciwdzia\u0142a\u0107 . Jego wiedza + intensywne zestawy \u0107wicze\u0144 , zaanga\u017cowanie i ch\u0119ci robi\u0105 \" cuda \" ! Otworzy\u0142 mi oczy , \u017ce to , to i to nie dzia\u0142a i nie funkcjonuje normalnie dlatego tak i tak si\u0119 dzieje . Teraz moja praca oraz treningi sta\u0142y si\u0119 w ko\u0144cu czyst\u0105 przyjemno\u015bci\u0105 a nie walk\u0105 z sam\u0105 sob\u0105 ! : ) POLECAM ! Na zako\u0144czenie dodam , \u017ce po kilku wizytach przyprowadzi\u0142a m swoj\u0105 mam\u0105 , kt\u00f3ra przez 9mc boryka\u0142a si\u0119 ostrym b\u00f3lem bark\u00f3w , odwiedzaj\u0105c przez ten okres przer\u00f3\u017cnych specjalist\u00f3w , doktor\u00f3w p\u0142ac\u0105c przy tym maj\u0105tek . W dodatku nikt przez ten czas nie pom\u00f3g\u0142 jej ani troch\u0119 i codziennie z b\u00f3lu \u0142yka\u0142a po 2 – 3 tabletki przeciw b\u00f3lowe . Po wizycie u Krzy\u015bka okaza\u0142o si\u0119 , \u017ce ma zerwane \u015bci\u0119gna . A po jego pierwszej , 3 godzinnej , bardzo bolesnej wizycie i masa\u017cach , przetrwa\u0142a do nast\u0119pnej wizyty tylko z siniakami , ale ju\u017c BEZ tabletek przeciwb\u00f3lowych ! : ) Krzysiek jest naprawd\u0119 GODNYM POLECENIA SPECJALIST\u0104 . Tym bardziej je\u015bli chcecie si\u0119 leczy\u0107 , naturalnie , bez zb\u0119dnych proszk\u00f3w i weso\u0142ej atmosferze ! : )", "label": "Pozytywny"}
```

Step 2

Design the prompt structure

~ — root@luna-0200: /workspace — ssh selene | ~ — root@luna-0198: /workspace/alpa/benchmark/cupy — ssh selene | ~ — ssh selene

{"taskname": "squad", "context": "Architecturally, the school has a Catholic character. Atop the Main Building's gold dome is a golden statue of the Virgin Mary. Immediatel
y in front of the Main Building and facing it, is a copper statue of Christ with arms upraised with the legend \"Venite Ad Me Omnes\". Next to the Main Building is the Basi
lica of the Sacred Heart. Immediately behind the basilica is the Grotto, a Marian place of prayer and reflection. It is a replica of the grotto at Lourdes, France where the
 Virgin Mary reputedly appeared to Saint Bernadette Soubirous in 1858. At the end of the main drive (and in a direct line that connects through 3 statues and the Gold Dome)
, is a simple, modern stone statue of Mary.", "question": "To whom did the Virgin Mary allegedly appear in 1858 in Lourdes France?", "answer": "Saint Bernadette Soubirous"}
{"taskname": "squad", "context": "Architecturally, the school has a Catholic character. Atop the Main Building's gold dome is a golden statue of the Virgin Mary. Immediatel
y in front of the Main Building and facing it, is a copper statue of Christ with arms upraised with the legend \"Venite Ad Me Omnes\". Next to the Main Building is the Basi
lica of the Sacred Heart. Immediately behind the basilica is the Grotto, a Marian place of prayer and reflection. It is a replica of the grotto at Lourdes, France where the
 Virgin Mary reputedly appeared to Saint Bernadette Soubirous in 1858. At the end of the main drive (and in a direct line that connects through 3 statues and the Gold Dome)
, is a simple, modern stone statue of Mary.", "question": "What is in front of the Notre Dame Main Building?", "answer": "a copper statue of Christ"}
{"taskname": "squad", "context": "Architecturally, the school has a Catholic character. Atop the Main Building's gold dome is a golden statue of the Virgin Mary. Immediatel
y in front of the Main Building and facing it, is a copper statue of Christ with arms upraised with the legend \"Venite Ad Me Omnes\". Next to the Main Building is the Basi
lica of the Sacred Heart. Immediately behind the basilica is the Grotto, a Marian place of prayer and reflection. It is a replica of the grotto at Lourdes, France where the
 Virgin Mary reputedly appeared to Saint Bernadette Soubirous in 1858. At the end of the main drive (and in a direct line that connects through 3 statues and the Gold Dome)
, is a simple, modern stone statue of Mary.", "question": "The Basilica of the Sacred heart at Notre Dame is beside to which structure?", "answer": "the Main Building"}
{"taskname": "squad", "context": "Architecturally, the school has a Catholic character. Atop the Main Building's gold dome is a golden statue of the Virgin Mary. Immediatel
y in front of the Main Building and facing it, is a copper statue of Christ with arms upraised with the legend \"Venite Ad Me Omnes\". Next to the Main Building is the Basi
lica of the Sacred Heart. Immediately behind the basilica is the Grotto, a Marian place of prayer and reflection. It is a replica of the grotto at Lourdes, France where the
 Virgin Mary reputedly appeared to Saint Bernadette Soubirous in 1858. At the end of the main drive (and in a direct line that connects through 3 statues and the Gold Dome)
, is a simple, modern stone statue of Mary.", "question": "What is the Grotto at Notre Dame?", "answer": "a Marian place of prayer and reflection"}
{"taskname": "squad", "context": "Architecturally, the school has a Catholic character. Atop the Main Building's gold dome is a golden statue of the Virgin Mary. Immediatel
y in front of the Main Building and facing it, is a copper statue of Christ with arms upraised with the legend \"Venite Ad Me Omnes\". Next to the Main Building is the Basi
lica of the Sacred Heart. Immediately behind the basilica is the Grotto, a Marian place of prayer and reflection. It is a replica of the grotto at Lourdes, France where the
 Virgin Mary reputedly appeared to Saint Bernadette Soubirous in 1858. At the end of the main drive (and in a direct line that connects through 3 statues and the Gold Dome)
, is a simple, modern stone statue of Mary.", "question": "What sits on top of the Main Building at Notre Dame?", "answer": "a golden statue of the Virgin Mary"}
{"taskname": "squad", "context": "As at most other universities, Notre Dame's students run a number of news media outlets. The nine student-run outlets include three newspa
pers, both a radio and television station, and several magazines and journals. Begun as a one-page journal in September 1876, the Scholastic magazine is issued twice monthl
y and claims to be the oldest continuous collegiate publication in the United States. The other magazine, The Juggler, is released twice a year and focuses on student liter
ature and artwork. The Dome yearbook is published annually. The newspapers have varying publication interests, with The Observer published daily and mainly reporting univer
sity and other news, and staffed by students from both Notre Dame and Saint Mary's College. Unlike Scholastic and The Dome, The Observer is an independent publication and d
oes not have a faculty advisor or any editorial oversight from the University. In 1987, when some students believed that The Observer began to show a conservative bias, a l
iberal newspaper, Common Sense was published. Likewise, in 2003, when other students believed that the paper showed a liberal bias, the conservative paper Irish Rover went
into production. Neither paper is published as often as The Observer; however, all three are distributed to all students. Finally, in Spring 2008 an undergraduate journal f
or political science research, Beyond Politics, made its debut.", "question": "When did the Scholastic Magazine of Notre dame begin publishing?", "answer": "September 1876"
}
{"taskname": "squad", "context": "As at most other universities, Notre Dame's students run a number of news media outlets. The nine student-run outlets include three newspa
pers, both a radio and television station, and several magazines and journals. Begun as a one-page journal in September 1876, the Scholastic magazine is issued twice monthl
y and claims to be the oldest continuous collegiate publication in the United States. The other magazine, The Juggler, is released twice a year and focuses on student liter
ature and artwork. The Dome yearbook is published annually. The newspapers have varying publication interests, with The Observer published daily and mainly reporting univer
sity and other news, and staffed by students from both Notre Dame and Saint Mary's College. Unlike Scholastic and The Dome, The Observer is an independent publication and d
oes not have a faculty advisor or any editorial oversight from the University. In 1987, when some students believed that The Observer began to show a conservative bias, a l
iberal newspaper, Common Sense was published. Likewise, in 2003, when other students believed that the paper showed a liberal bias, the conservative paper Irish Rover went
into production. Neither paper is published as often as The Observer; however, all three are distributed to all students. Finally, in Spring 2008 an undergraduate journal f
or political science research, Beyond Politics, made its debut.", "question": "How often is Notre Dame's the Juggler published?", "answer": "twice"}
{"taskname": "squad", "context": "As at most other universities, Notre Dame's students run a number of news media outlets. The nine student-run outlets include three newspa
pers, both a radio and television station, and several magazines and journals. Begun as a one-page journal in September 1876, the Scholastic magazine is issued twice monthl

NVIDIA

Step 2

Design the prompt structure

```
  task_templates: # task_templates for all existing_tasks and new_tasks are required.
  - taskname: "squad" # The task name
    prompt_template: "<|VIRTUAL_PROMPT_0|>Context: {context} Question: {question} Answer: {answer}" # Prompt template for task, specify
 virtual prompt positions with <|VIRTUAL_PROMPT_#|>
    total_virtual_tokens: 10 # Sum of tokens in virtual_token_splits must add to this number. Can differ between new and existing tasks
, but must match across all new tasks being tuned at the same time.
    virtual_token_splits: [10] # number of virtual tokens to be inserted at each VIRTUAL PROMPT location, must add to total_virtual_tok
ens
    truncate_field: "context" # The {field} in the prompt template whose text will be truncated if the input is too long, if null, inpu
ts that are too long will just be skipped.
    answer_field: "answer" # Answer/Target field
    answer_only_loss: True # If true, the loss will only be calculated with answer_field text vs. ground truth. If false, the loss will
 be calculated over entire sentence.
```

```
  task_templates: # task_templates for all existing_tasks and new_tasks are required.
  - taskname: "squad" # The task name
    prompt_template: "<|VIRTUAL_PROMPT_0|>{context}<|VIRTUAL_PROMPT_1|>{question}<|VIRTUAL_PROMPT_2|>{answer}" # Prompt template for ta
sk, specify virtual prompt positions with <|VIRTUAL_PROMPT_#|>
    total_virtual_tokens: 100 # Sum of tokens in virtual_token_splits must add to this number. Can differ between new and existing task
s, but must match across all new tasks being tuned at the same time.
    virtual_token_splits: [80,15,5] # number of virtual tokens to be inserted at each VIRTUAL PROMPT location, must add to total_virtua
l_tokens
    truncate_field: "context" # The {field} in the prompt template whose text will be truncated if the input is too long, if null, inpu
ts that are too long will just be skipped.
    answer_field: "answer" # Answer/Target field
    answer_only_loss: True # If true, the loss will only be calculated with answer_field text vs. ground truth. If false, the loss will
 be calculated over entire sentence.
```

Step 2

Alternative examples

```
config.model.task_templates = [
    {
        "taskname": "sentiment",
        "prompt_template": "<|VIRTUAL_PROMPT_0|> {sentence} sentyment:{label}",
        "total_virtual_tokens": 10,
        "virtual_token_splits": [10],
        "truncate_field": None,
        "answer_only_loss": True,
        "answer_field": "label",
    },]
```

```
task_templates: # Add more/replace tasks as needed, these are just examples
- taskname: "boolq" # The task name
  prompt_template: "<|VIRTUAL_PROMPT_0|> Passage: {passage} <|VIRTUAL_PROMPT_1|> \nQuestion: {question} \nAnswer: {answer}" # Prompt template for task, specify virtual prompt positions with <|VIRTUAL_PROMPT_#
|>
  total_virtual_tokens: 30 # Sum of tokens in virtual_token_splits must add to this number. Can differ between new and existing tasks, but must match across all new tasks being tuned at the same time.
  virtual_token_splits: [20, 10] # number of virtual tokens to be inserted at each VIRTUAL PROMPT location, must add to total_virtual_tokens
  truncate_field: "passage" # The {field} in the prompt template whose text will be truncated if the input is too long, if null, inputs that are too long will just be skipped.
  answer_only_loss: True
  answer_field: "answer"

- taskname: "intent_and_slot"
  prompt_template: "<|VIRTUAL_PROMPT_0|> intent options: {intent_options} <|VIRTUAL_PROMPT_1|> slot options: {slot_options} <|VIRTUAL_PROMPT_2|> {utterance} \nintent: {intent} \nslot: {slot}"
  total_virtual_tokens: 30
  answer_only_loss: False
  virtual_token_splits: [15, 10, 5]
  truncate_field: null

- taskname: "rte"
  prompt_template: "<|VIRTUAL_PROMPT_0|>{premise}\n{hypothesis}\nAnswer: {answer}"
  total_virtual_tokens: 9
  virtual_token_splits: [9]
  truncate_field: null
  answer_only_loss: True
  answer_field: "answer"
```

Step 3

Training configuration

~ — root@luna-0200: /workspace — ssh selene | ~ — root@luna-0198: /workspace/alpa/benchmark/cupy — ssh • ssh draco-rnom3 | ~ — ssh selene

```
  checkpoint_callback_params:
    monitor: val_loss
    save_top_k: 5
    mode: min
    save_nemo_on_train_end: False
    filename: "megatron_gpt_prompt_learn--{val_loss:.3f}-{step}"
    model_parallel_size: ${prompt_learning.model.model_parallel_size}
    save_best_model: True

model:
  seed: 1234
  nemo_path: ${prompt_learning.run.results_dir}/results/megatron_gpt_prompt.nemo # the place to save prompt learning nemo checkpoint
  virtual_prompt_style: 'p-tuning' # One of 'p-tuning', 'prompt-tuning', or 'inference'. We recommend 'p-tuning' over 'prompt-tuning'.
  tensor_model_parallel_size: 1
  pipeline_model_parallel_size: 1
  model_parallel_size: ${multiply:${.tensor_model_parallel_size}, ${.pipeline_model_parallel_size}}
  encoder_seq_length: 2048
  global_batch_size: 64
  micro_batch_size: 8

  restore_path: null # used to restore from a prompt tuned checkpoint and add new tasks
  language_model_path: /lustre/fsw/sa/adamg/nemogpt/gpt5b/nemo_gpt5B_fp16_tp2.nemo
  #  language_model_path: ${prompt_learning.run.convert_dir}/results/megatron_gpt.nemo # Restore lanugage model from pre-trained .nemo checkpoint
  existing_tasks: [] # if restore from a prompt tuned checkpoint and add new tasks, existing task names should be included here.
  new_tasks: ["squad"] # multiple tasks can be tuned at the same time

  task_templates: # task_templates for all existing_tasks and new_tasks are required.
  - taskname: "squad" # The task name
    prompt_template: "<|VIRTUAL_PROMPT_0|>Context: {context} Question: {question} Answer: {answer}" # Prompt template for task, specify virtual prompt positions with <|VIRT
UAL_PROMPT_#|>
    total_virtual_tokens: 10 # Sum of tokens in virtual_token_splits must add to this number. Can differ between new and existing tasks, but must match across all new tasks
 being tuned at the same time.
    virtual_token_splits: [10] # number of virtual tokens to be inserted at each VIRTUAL PROMPT location, must add to total_virtual_tokens
    truncate_field: "context" # The {field} in the prompt template whose text will be truncated if the input is too long, if null, inputs that are too long will just be ski
pped.
    answer_field: "answer" # Answer/Target field
    answer_only_loss: True # If true, the loss will only be calculated with answer_field text vs. ground truth. If false, the loss will be calculated over entire sentence.

  prompt_learning: # Prompt tunin specific params
    new_prompt_init_methods: null # e.g ['text'], List of 'text' or 'random', should correspond to tasks listed in new tasks
    new_prompt_init_text: null # e.g ['some init text goes here'], some init text if init method is text, or None if init method is random

  p_tuning: # P-tuning specific params
    dropout: 0.0
    num_layers: 2

  data:
    train_ds:
      - ${data_dir}/prompt_data/v1.1/squad_train.jsonl # multiple prompt dataset can be given at the same time
                                                                                                                                78,11         63%
```

NVIDIA

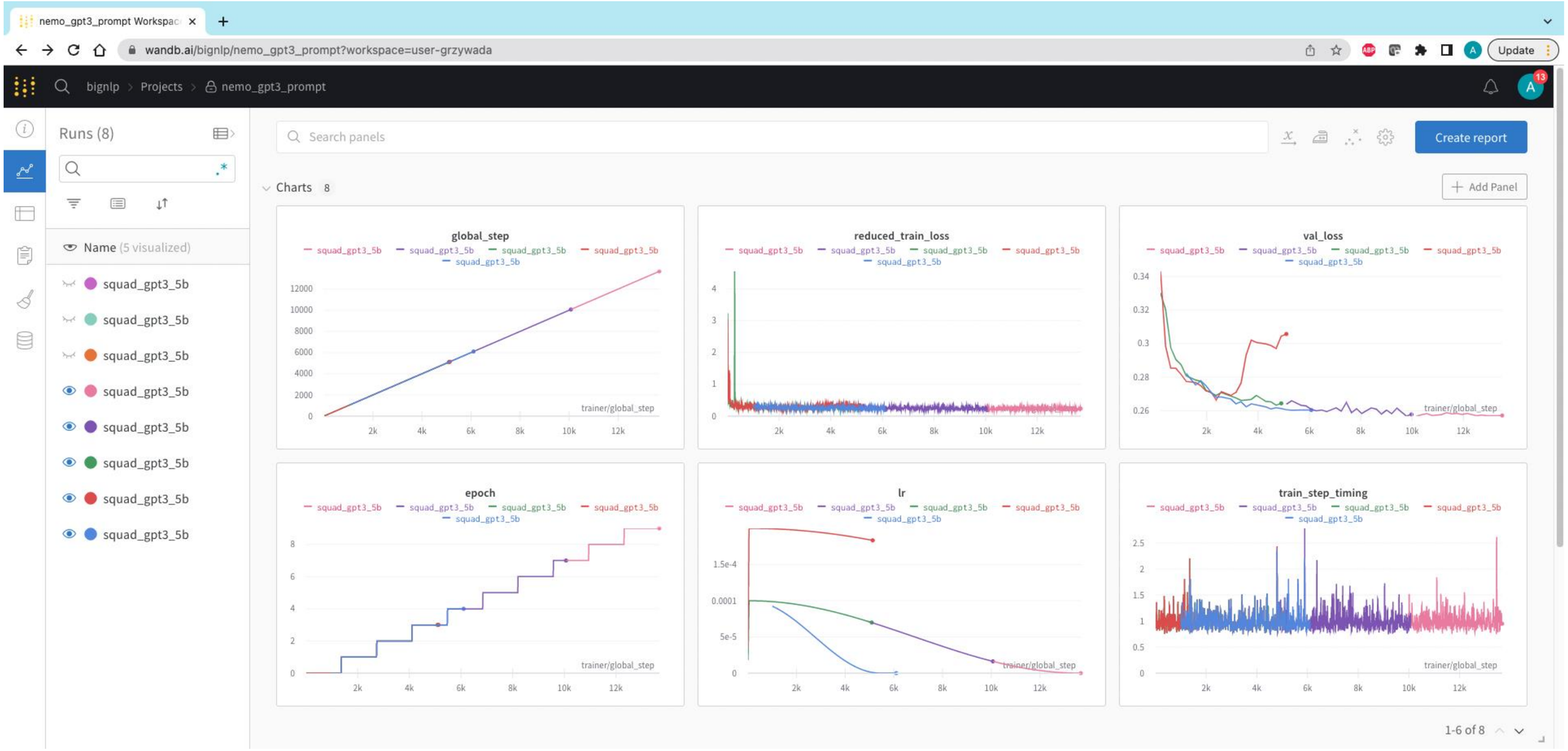
Step 4

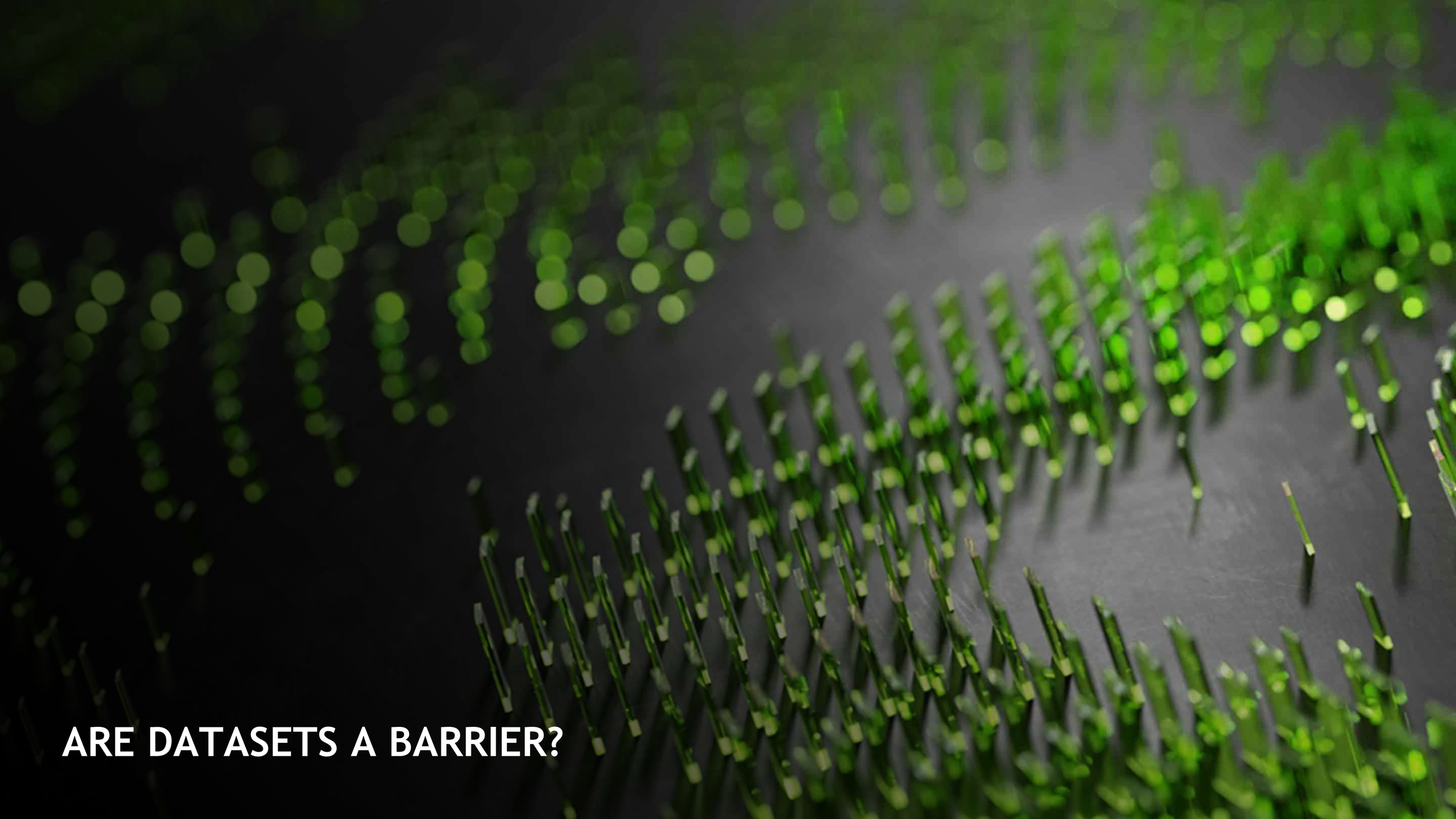
Kick off the training process

```
adamg@selene-login-01:~/bignlp/workspace/bignlp-scripts/conf/prompt_learning/gpt3$ ls
squad.yaml
adamg@selene-login-01:~/bignlp/workspace/bignlp-scripts/conf/prompt_learning/gpt3$ vim squad.yaml
adamg@selene-login-01:~/bignlp/workspace/bignlp-scripts/conf/prompt_learning/gpt3$ cd ..
adamg@selene-login-01:~/bignlp/workspace/bignlp-scripts/conf/prompt_learning$ cd ..
adamg@selene-login-01:~/bignlp/workspace/bignlp-scripts/conf$ cd ..
adamg@selene-login-01:~/bignlp/workspace/bignlp-scripts$ python3 main.py
```

Step 5

Monitor the training process




ARE DATASETS A BARRIER?

Unsupervised models

Limited data processing complexity

The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only

The Falcon LLM team

Guilherme Penedo [1] Quentin Malartic [2]
Daniel Hesslow [1] Ruxandra Cojocaru [2] Alessandro Cappelli [1] Hamza Alobeidli [2] Baptiste Pannier [1]
Ebtesam Almazrouei [2] Julien Launay [1 3]

https://huggingface.co/datasets/tiiuae/falcon-refinedweb

Abstract

Large language models are commonly trained on a mixture of filtered web data and curated "high-quality" corpora, such as social media conversations, books, or technical papers. This curation process is believed to be necessary to produce performant models with broad zero-shot generalization abilities. However, as larger models requiring pretraining on trillions of tokens are considered, it is unclear how scalable is curation and whether we will run out of unique high-quality data soon. At variance with previous beliefs, we show that properly filtered and deduplicated web data alone can lead to powerful models; even significantly outperforming models from the state-of-the-art trained on The Pile. Despite extensive filtering, the high-quality data we extract from the web is still plentiful, and we are able to obtain five trillion tokens from CommonCrawl. We publicly release an extract of 600 billion tokens from our REFINEDWEB dataset, and 1.3/7.5B parameters language models trained on it*.

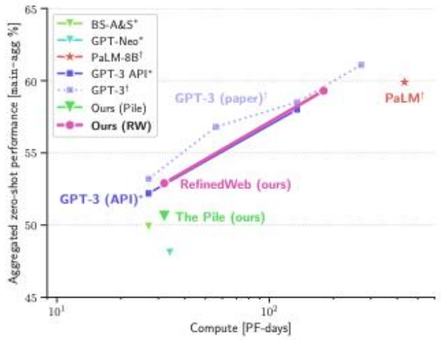
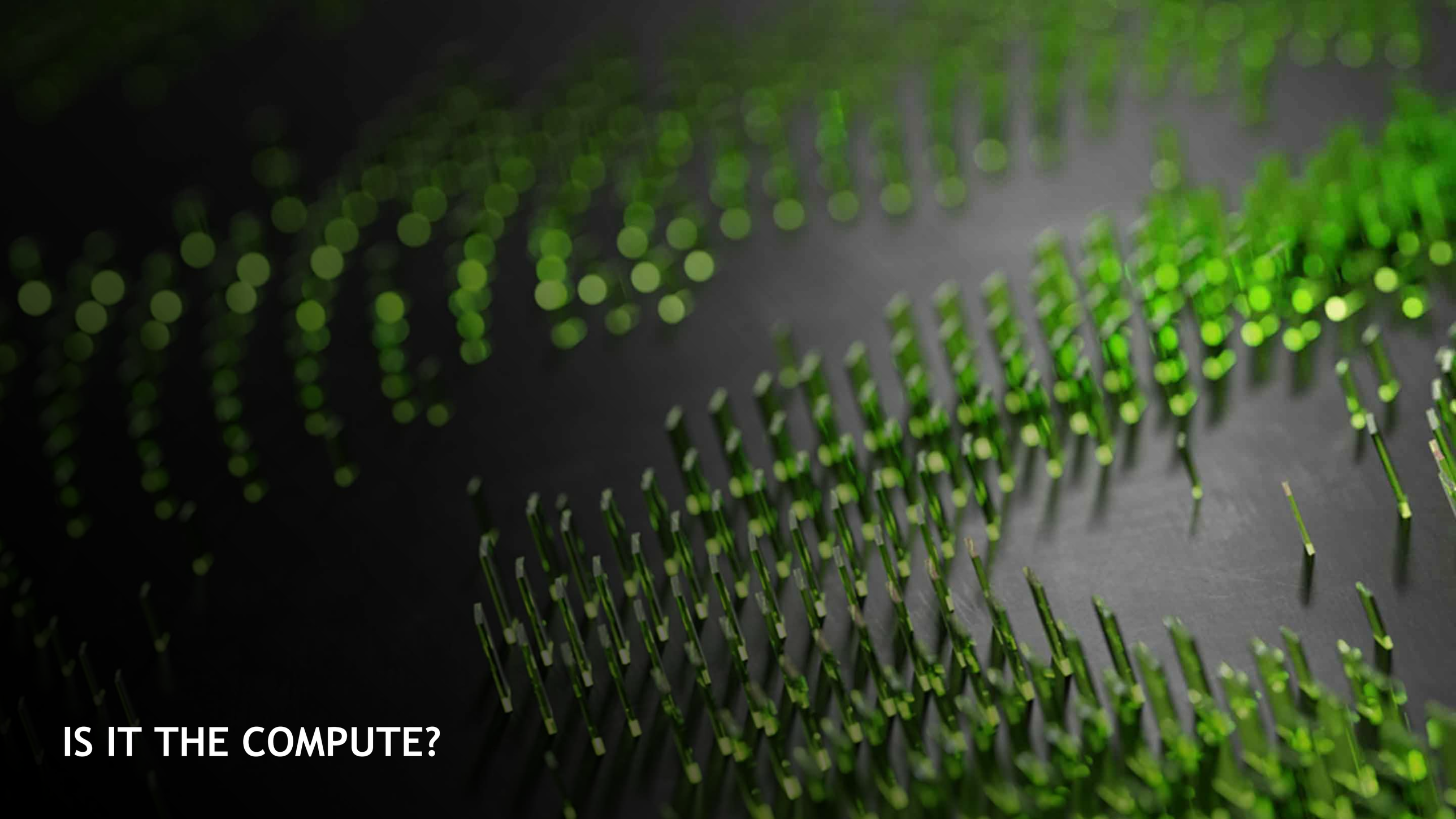


Figure 1. **Models trained on ●REFINEDWEB alone outperform models trained on curated corpora.** Zero-shot performance on our `main-agg` task aggregate (see Section 4.1 for details). At equivalent compute budgets, our models significantly outperform publicly available models trained on ▼ The Pile, and match the performance of the ■ GPT-3 models when tested within our evaluation setup.

[1]LightOn [2]Technology Innovation Institute, 9639 Masdar City, Abu Dhabi, United Arab Emirates [3]LPENS, École normale supérieure. Contact: <falconllm@tii.ae>.

*Details about how to access Falcon LLM open source is available on falconllm.tii.ae


IS IT THE COMPUTE?

Training LLMs is computationally intensive

GPT-3 Training Time on NVIDIA A100 GPUs

	Time to train 300B tokens in days (A100) – BF16			
	800 GPUs (5x DGX SuperPod)	**480 GPUs (3x DGX SuperPod)**	**160 GPUs (1x DGX SuperPod)**	**64 GPUs (8x DGX A100)**
GPT-3: 126M	0.07	0.12	0.37	0.92
GPT-3: 5B	0.8	1.3	3.9	9.8
GPT-3: 20B	3.6	6	18.1	45.3
GPT-3: 40B	6.6	10.9	32.8	82
GPT-3: 175B	28	46.7	140	349.9

LLAMA 2 TRAINING TIME

Hypothetical Training Time on single NVIDIA A100 GPUs

Single GPU

meta-llama

24 years +

LLAMA 2 TRAINING TIME

Training Time on NVIDIA A100 GPUs

DiRAC: Tursa

meta-llama

157 days

ABOUT ME

Adam Grzywaczewski – adamg@nvidia.com