

A Practical Perspective on LLMs for Enterprises



Rajiv Shah



@rajistics

raj@huggingface.co



VIDEO LINK FOR TALK: <https://youtu.be/dKBD-3hnjW0>

Applications

Question

A llama is in my lawn. How do I get rid of him?

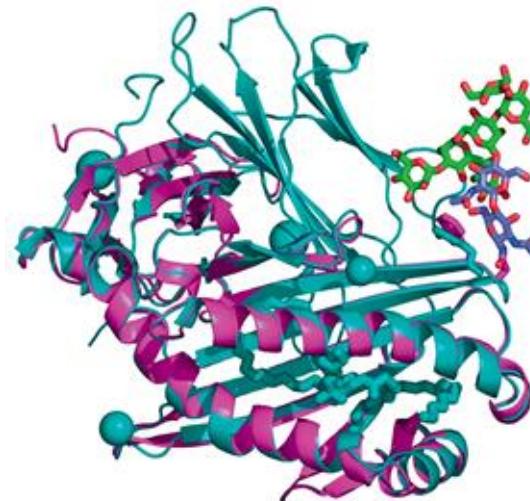
Answer

There are a few options, depending on the size and temperament of your pet.

1. Put it into a cage/carrier until you can find someone to take care of it.
2. Call an animal shelter or veterinarian for advice on how best to contain and feed such a large creature (probably not feasible).
3. Use force.



Rajiv Shah @rajistics



JS draw_scatterplot.js draw_scatterplot.py draw_scatterplot.r

```

1 import matplotlib.pyplot as plt
2
3 def draw_scatterplot(x_values, y_values):
4     plt.scatter(x_values, y_values, s=20)
5     plt.title("Scatter Plot")
6     plt.xlabel("x values")
7     plt.ylabel("y values")
8     plt.show()

```

Copilot

My Joke website

[really funny joke 1] [push to reveal punchline]

[some, but joke 2] [push to reveal punchline]

© OpenAI 2023

My Joke Website

Joke 1: Why don't scientists trust atoms

Reveal Punchline

Because they make up everything!

It's a skeleton's least favorite room in

Enterprise AI



When should we use LLMs?

Which LLM?



Rajiv Shah @rajistics

VIDEO LINK FOR TALK: <https://youtu.be/dKBD-3hnjW0>



Analytical



Transformative

What is your ChatGPT Strategy?

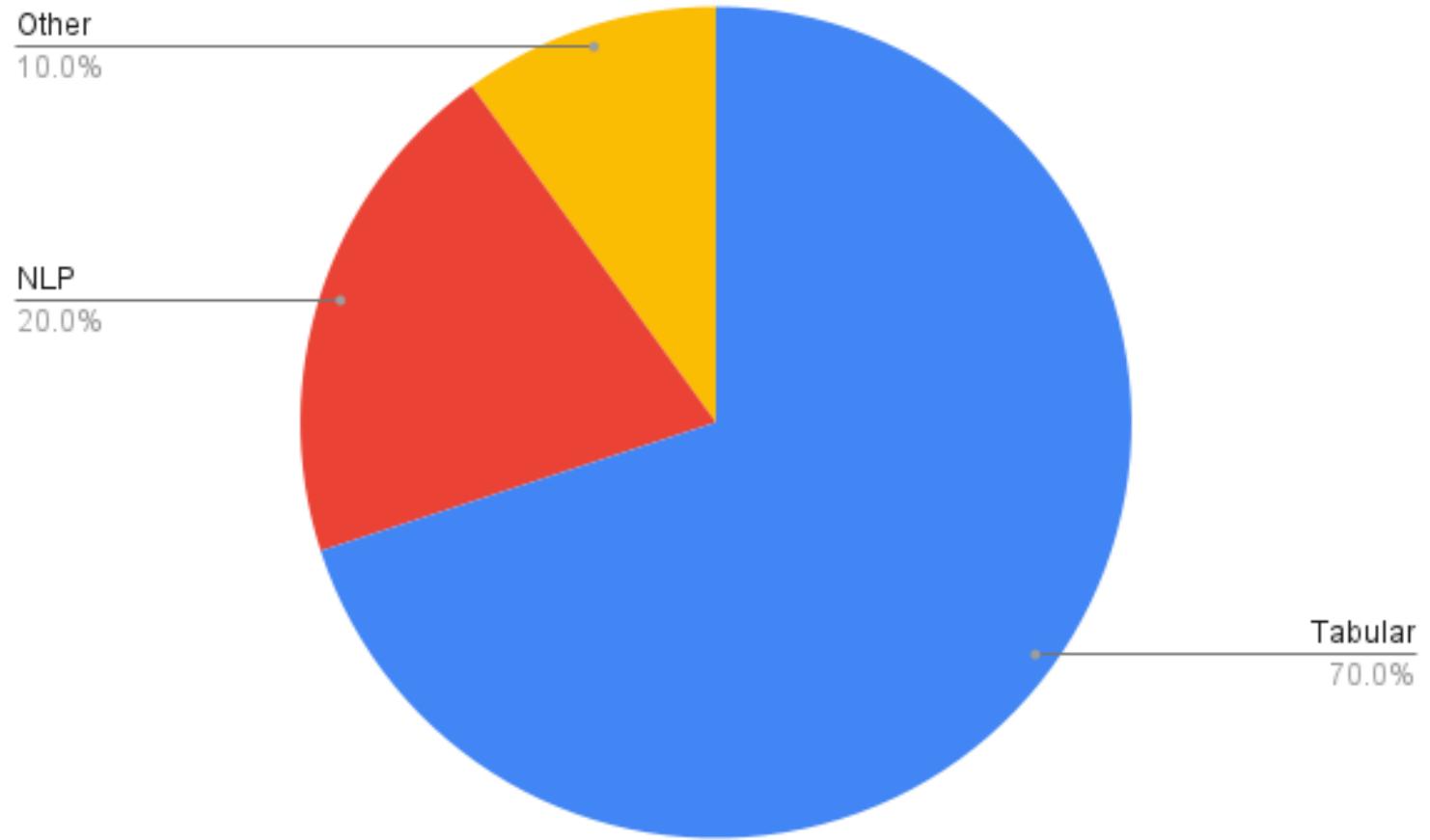


Rajiv Shah @rajis

ChatGPT affects only a subset

In the near term:

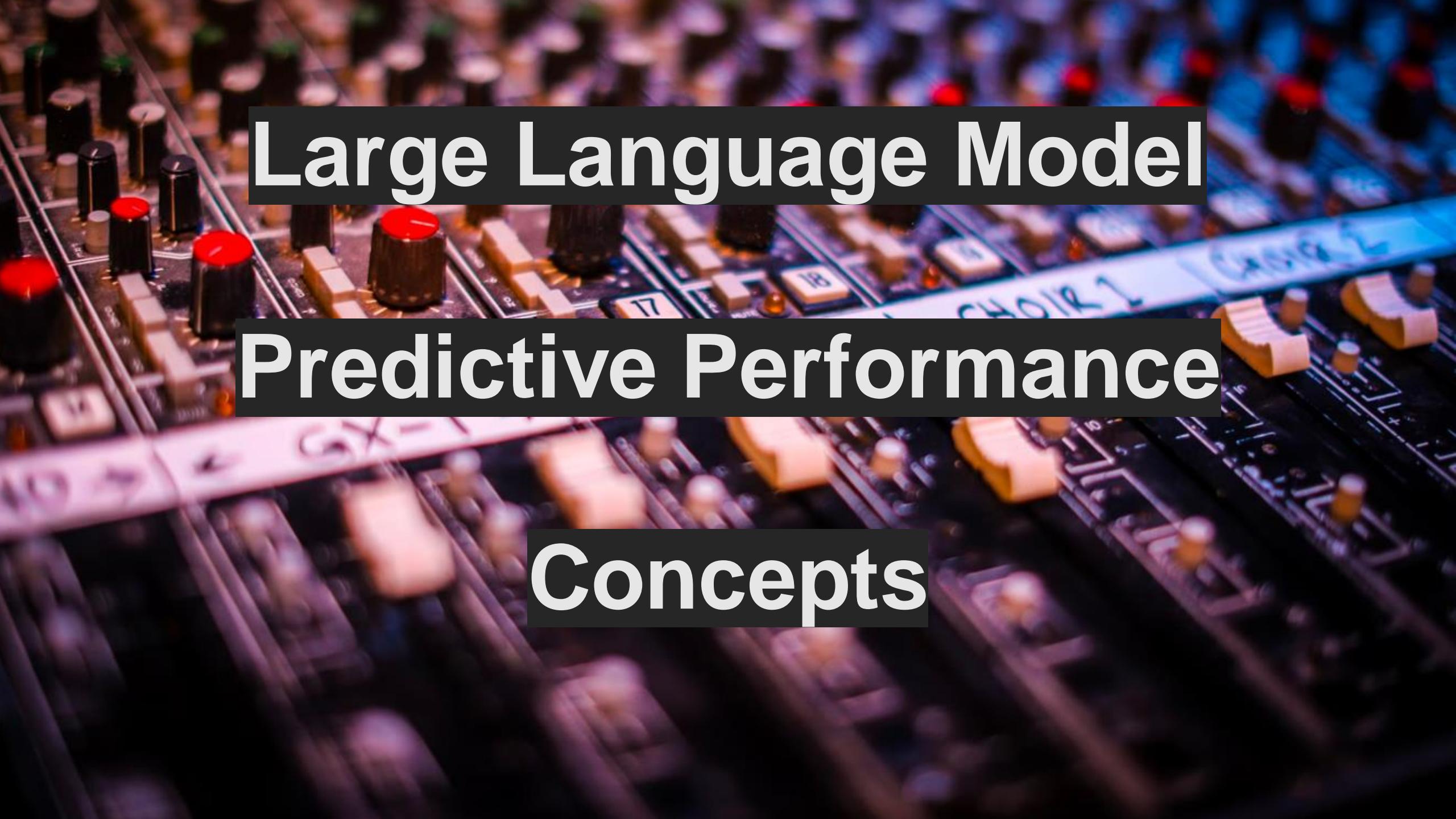
Still have to do churn,
forecast models and
dashboards



Rajiv Shah @rajistics

Source: Raj's Experience





Large Language Model

Predictive Performance

Concepts

GPT-2



Write With Transformer `gpt2` ⓘ

🔀 Shuffle initial text

⬇ Trigger autocomplete or `tab`

Select suggestion `↑` `↓` and `enter`

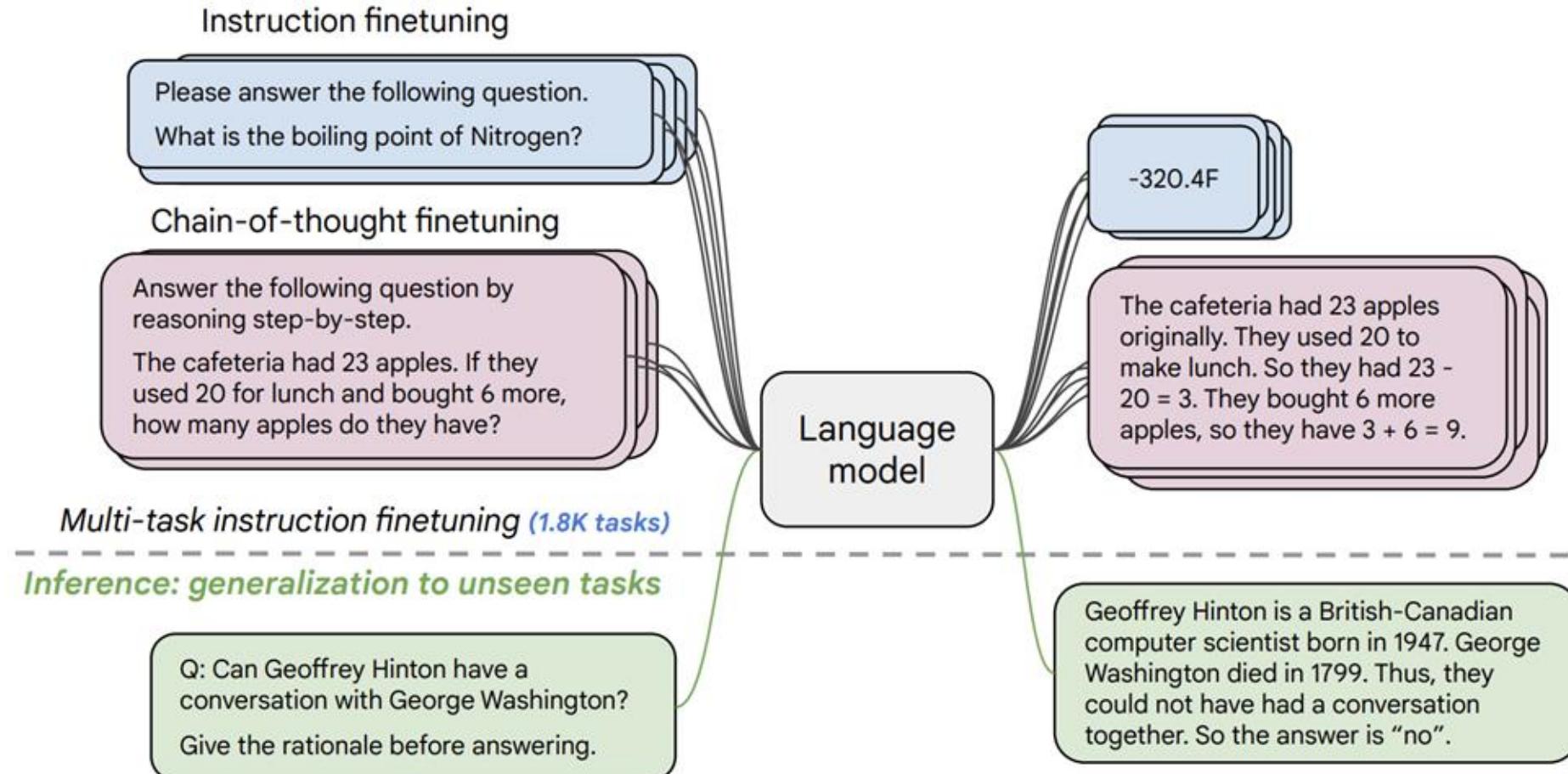
Cancel suggestion `esc`

Save & P



Rajiv Shah @

Fine Tuning: Instruction Tuning



Zero Shot Learning

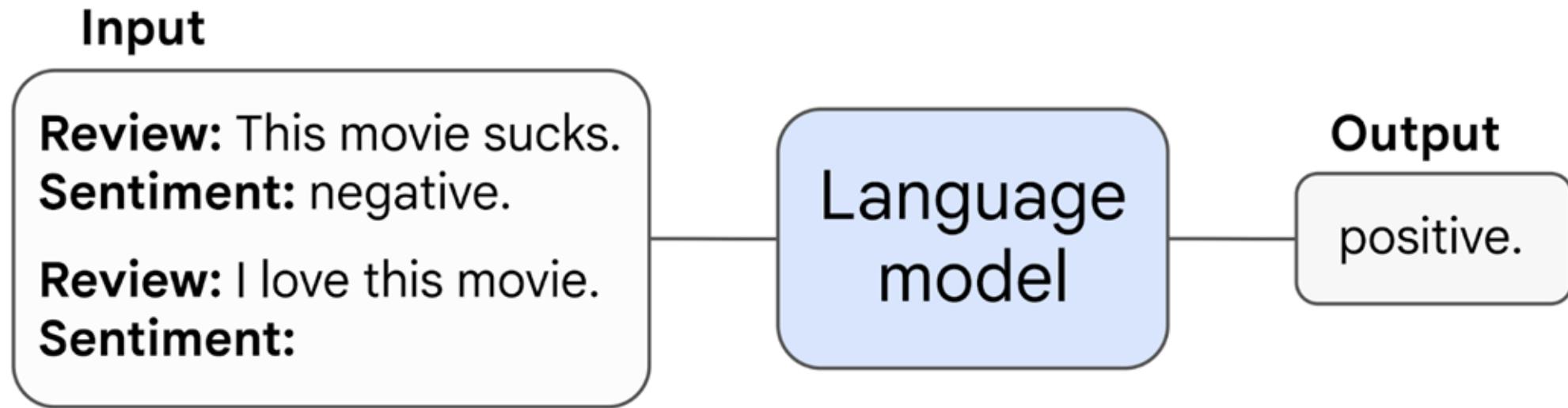
LLMs can now solve
various tasks

Task Description:
Convert English to French

Prompt:
cheese =>



Few-Shot Learning



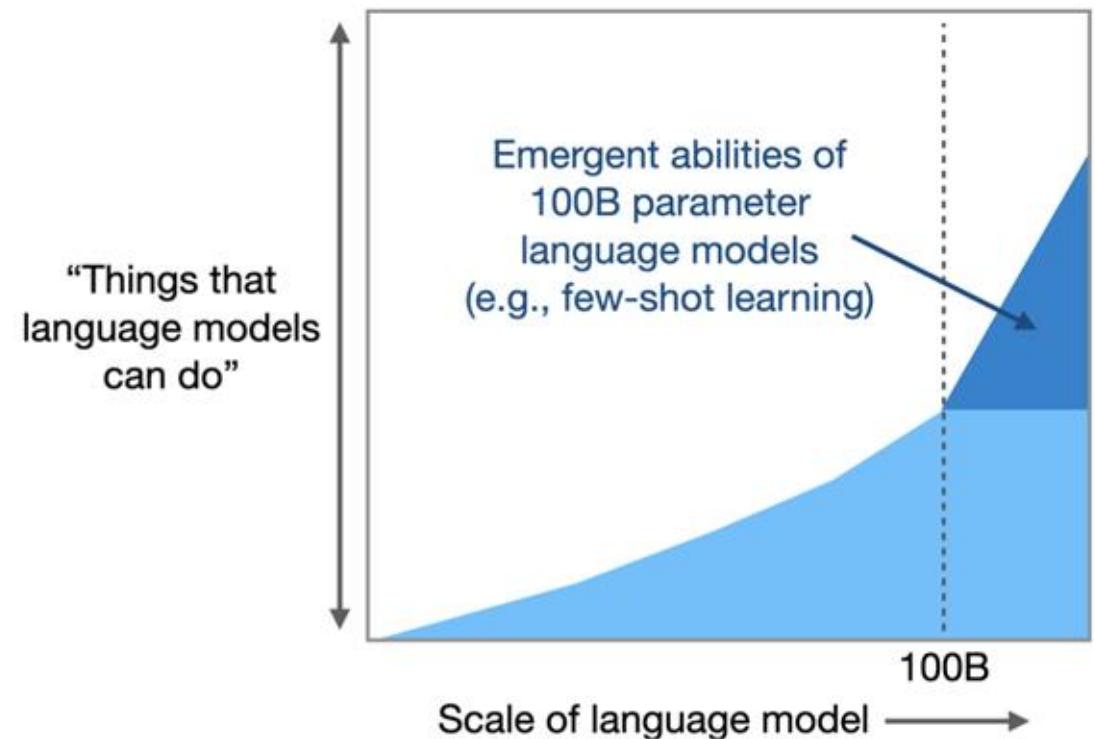
So many use cases!

Q&A Answer questions based on existing knowl...	Grammar correction Corrects sentences into standard English.	Spreadsheet creator Create spreadsheets of various kinds of dat...	JavaScript helper chatbot Message-style bot that answers JavaScript ...
Summarize for a 2nd grader Translates difficult text into simpler concep...	Natural language to OpenAI API Create code to call to the OpenAI API usin...	ML/AI language model tutor Bot that answers questions about language...	Science fiction book list maker Create a list of items for a given topic.
Text to command Translate text into programmatic commands.	English to other languages Translates English text into French, Spanish...	Tweet classifier Basic sentiment detection for a piece of text.	Airport code extractor Extract airport codes from text.
Natural language to Stripe API Create code to call the Stripe API using nat...	SQL translate Translate natural language to SQL queries.	SQL request Create simple SQL queries.	Extract contact information Extract contact information from a block of ...
Parse unstructured data Create tables from long form text	Classification Classify items into categories via example.	JavaScript to Python Convert simple JavaScript expressions into ...	Friend chat Emulate a text message conversation.
Python to natural language Explain a piece of Python code in human un...	Movie to Emoji Convert movie titles into emoji.	Mood to color Turn a text description into a color.	Write a Python docstring An example of how to create a docstring for ...
Calculate Time Complexity Find the time complexity of a function.	Translate programming languages Translate from one programming language ...	Analogy maker Create analogies. Modified from a communi...	JavaScript one line function Turn a JavaScript function into a one liner.
Advanced tweet classifier Advanced sentiment detection for a piece o...	Explain code Explain a complicated piece of code.	Micro horror story creator Creates two to three sentence short horror ...	Third-person converter Converts first-person POV to the third-pers...
Keywords Extract keywords from a block of text.	Factual answering Guide the model towards factual answering ...	Notes to summary Turn meeting notes into a summary.	VR fitness idea generator Create ideas for fitness and virtual reality g...
Ad from product description Turn a product description into ad copy.	Product name generator Create product names from examples word...	ESRB rating Categorize text based upon ESRB ratings.	Essay outline Generate an outline for a research topic.
TL;DR summarization Summarize text by adding a 'tl;dr' to the en...	Python bug fixer Find and fix bugs in source code.	Recipe creator (eat at your own risk) Create a recipe from a list of ingredients.	Chat Open ended conversation with an AI assist...



In Context Learning

LLMs are intriguing
because of the **emergence**
of In-Context Learning



Rajiv Shah @rajistics

Source: Jason Wei

Fine-Tuning: Reinforcement Learning Human Feedback (RLHF)

Step 2

Collect comparison data, and train a reward model.

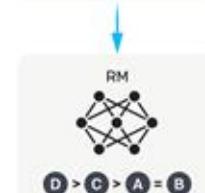
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.



The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



Fine-Tuning: RLHF

Ranking outputs

To be ranked

B A team of researchers from Yale University and University of California, Davis studied the vocalization patterns of several different types of parrots. They found that parrots like to mimic human speech, and can produce a wide range of sounds, such as whistles, squawks, and other types of vocalizations...

Rank 1 (best)

A A research group in the United States has found that parrots can imitate human speech with ease, and some of them can even do so in the same way as humans. This group studied the sounds that parrots make in their natural habitats and found that they use their tongues and beaks in ways that are strikingly...

C Parrots have been found to have the ability to understand numbers. Researchers have found that parrots can understand numbers up to six. In a series of experiments, the parrots were able to identify the amount of food items under a number of cups...

Rank 2

E Scientists have found that green-winged parrots can tell the difference between two noises that are the same except for the order in which they are heard. This is important because green-winged parrots are known to imitate sounds. This research shows that they are able to understand the difference between sounds.

D Current research suggests that parrots see and hear things in a different way than humans do. While humans see a rainbow of colors, parrots only see shades of red and green. Parrots can also see ultraviolet light, which is invisible to humans. Many birds have this ability to see ultraviolet light, an ability

Rank 4

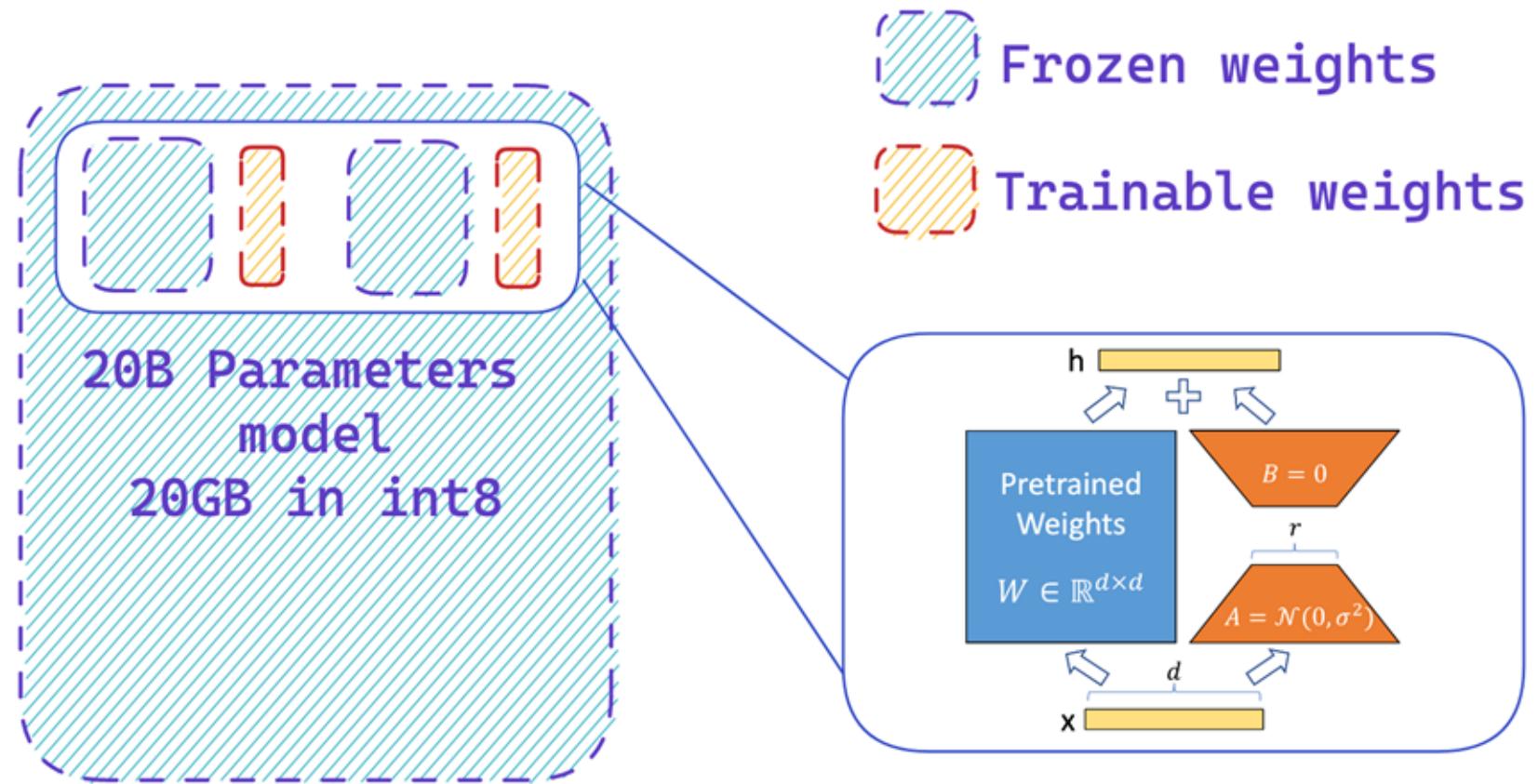
Rank 5 (worst)



Fine Tuning:

Parameter-Efficient Fine-Tuning (PEFT)

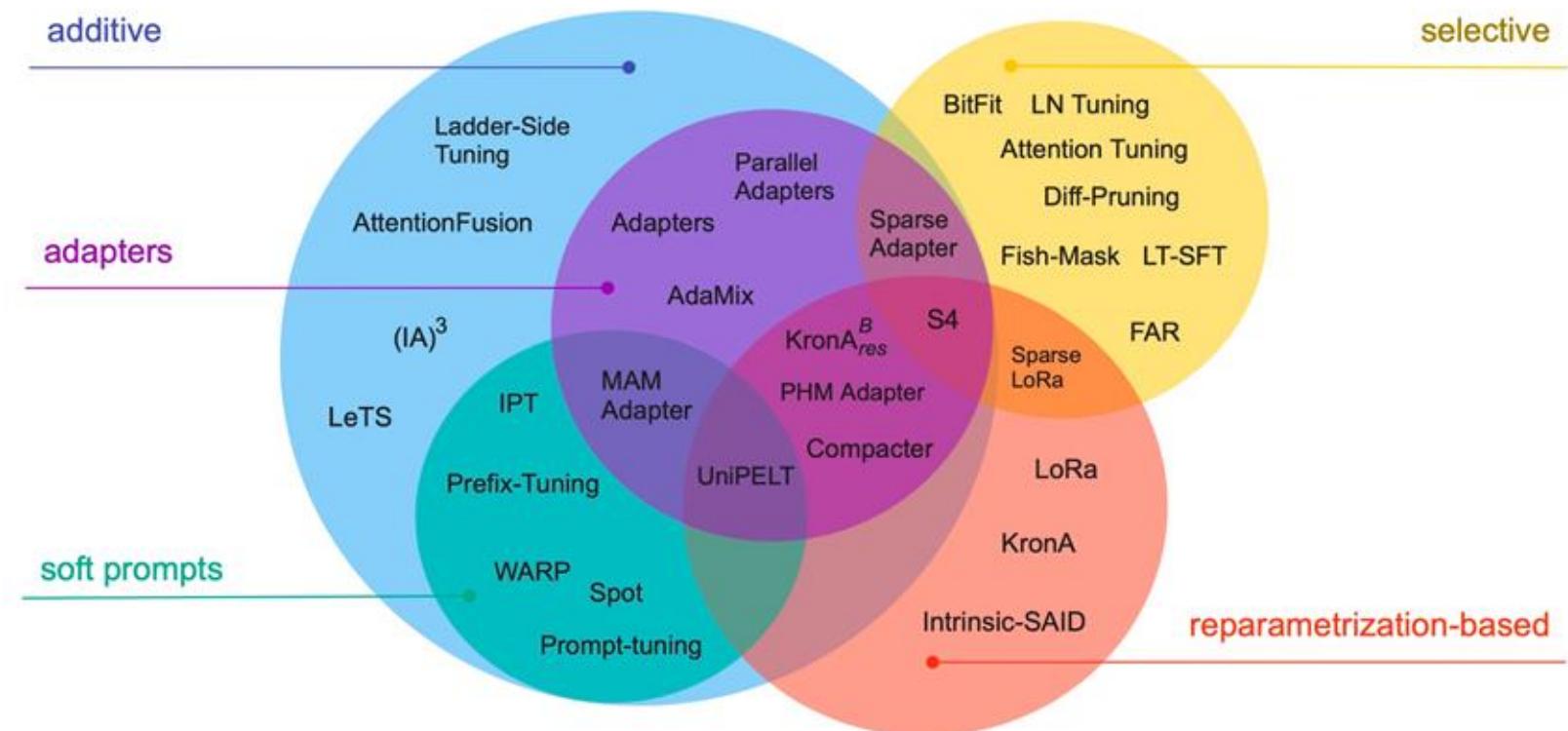
- Fine tuning a subset of LLM parameters



Fine Tuning:

Parameter-Efficient Fine-Tuning (PEFT)

- Fine tuning a subset of LLM parameters

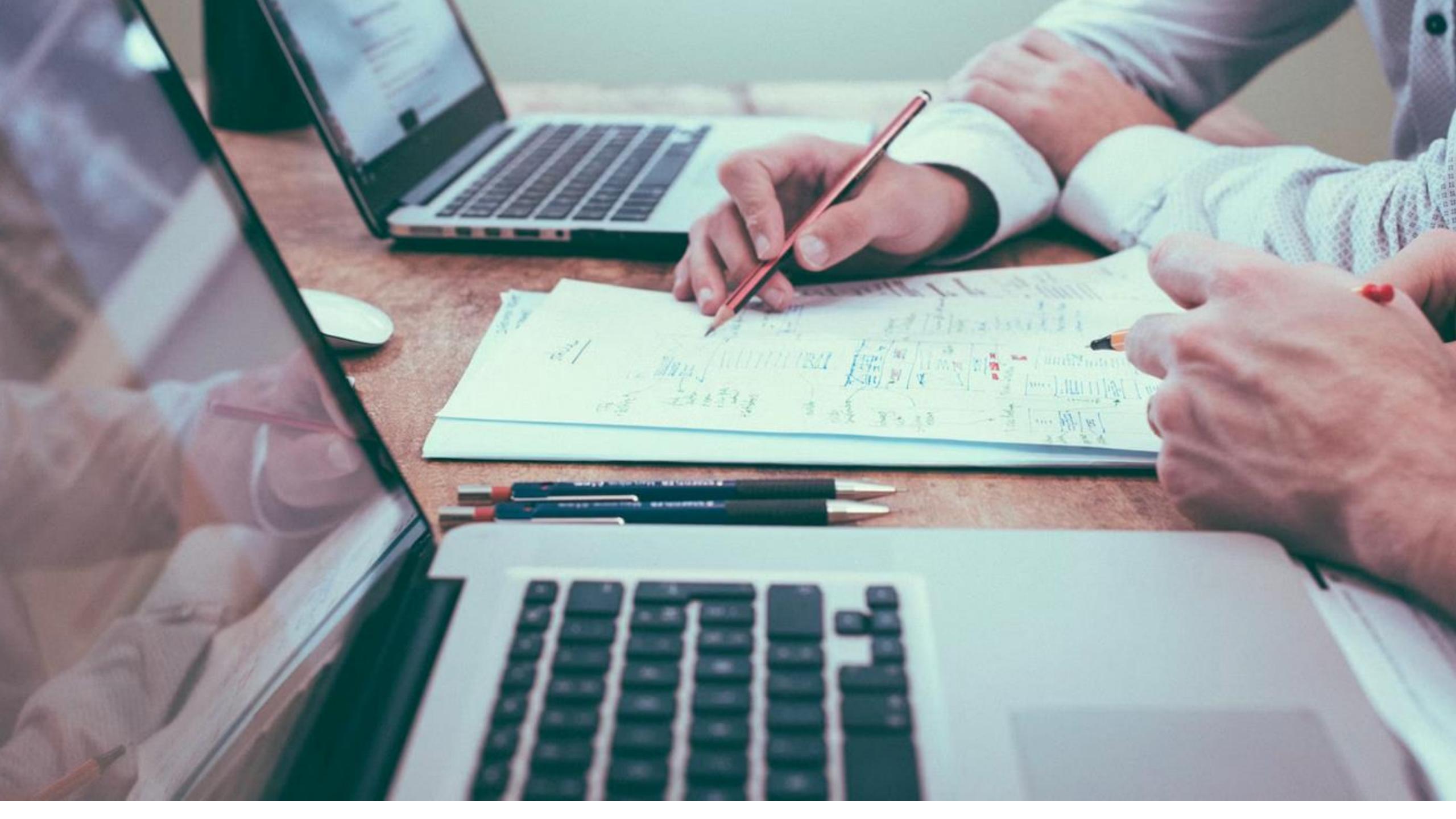


Strategies for Better LLM Predictions



- Many types of Fine Tuning
 - Instruction Tuning
 - RLHF
 - PEFT





When should we use LLMs?

LLMs: GPT-3,
FLAN-T5

Models:
BERT



Text Classification

Given a news article, classify its topic.

Possible labels: 1. World 2. Sports 3. Business 4. Sci/Tech

Article: A nearby star thought to harbor comets and asteroids now appears to be home to planets, too.

Label: Sci/Tech

Article: Soaring crude prices plus worries about the economy and the outlook for earnings are expected to hang over the stock market next week during the depth of the summer doldrums.

Label: Business

Article: Murtagh a stickler for success Northeastern field hockey coach Cheryl Murtagh doesn't want the glare of the spotlight that shines on her to detract from a team that has been the America East champion for the past three years and has been to the NCAA tournament 13 times.

Label:

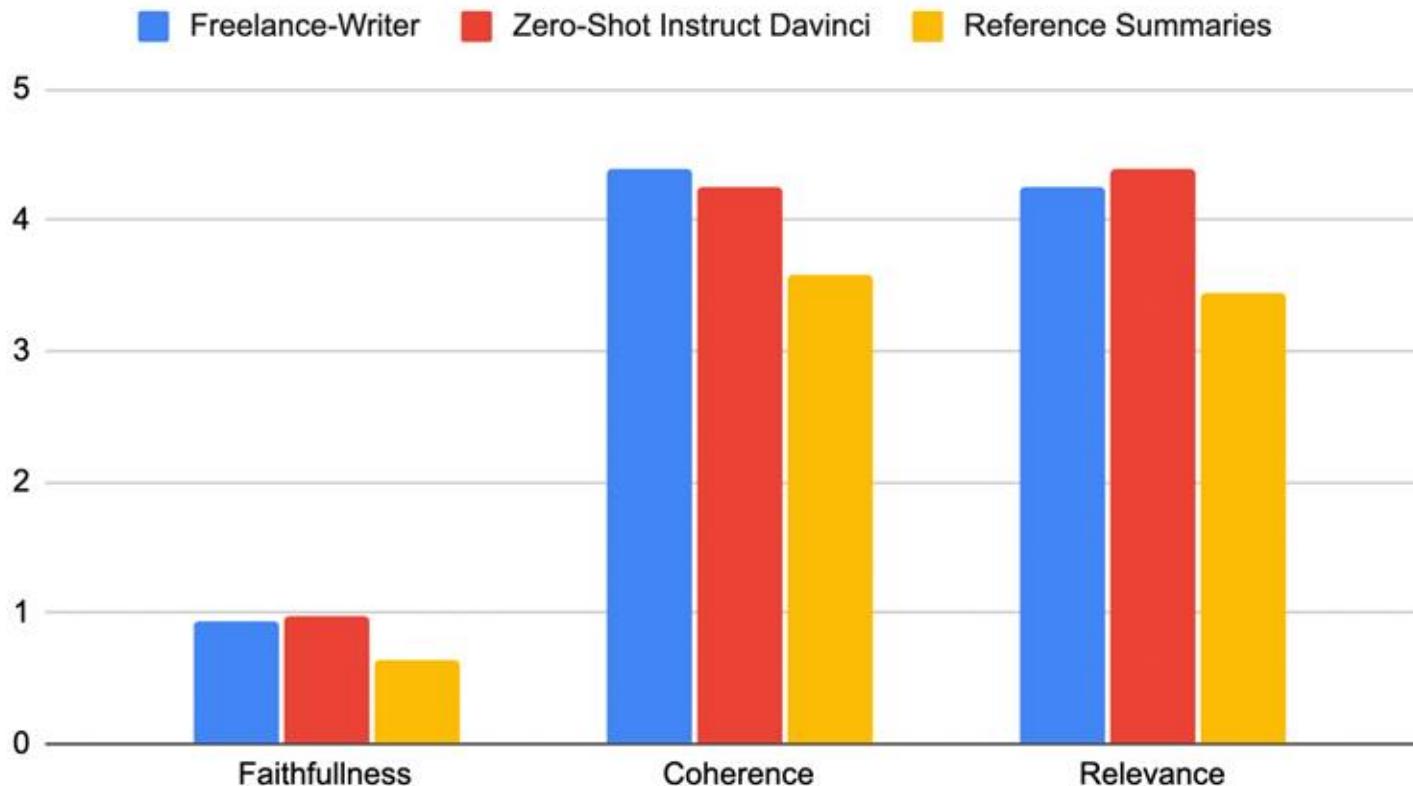


Rajiv Shah @rajistics

VIDEO LINK FOR TALK: <https://youtu.be/dKBD-3hnjW0>

Summarization

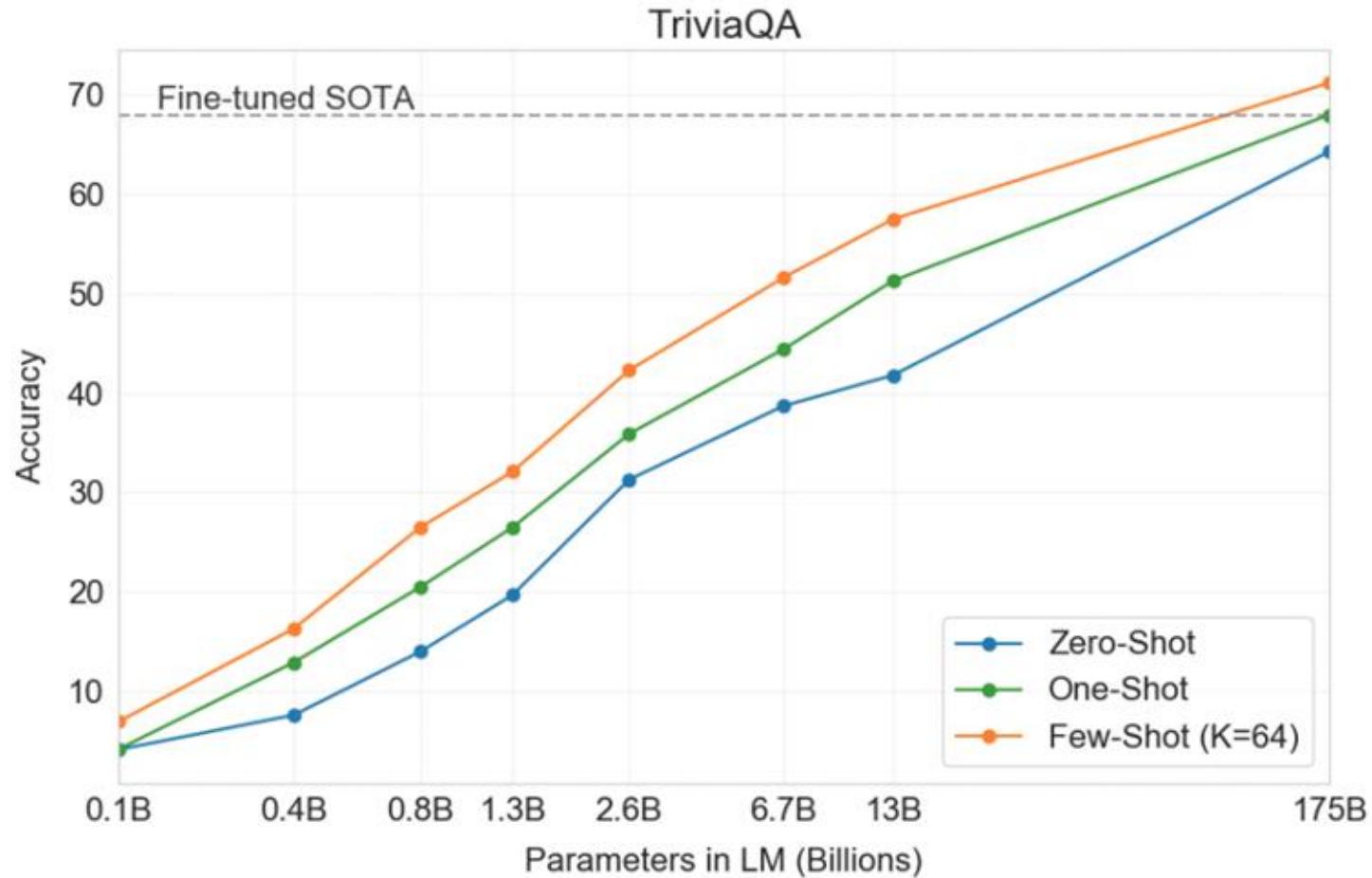
Freelance writers versus GPT-3



Rajiv Shah @rajistics

<https://arxiv.org/pdf/2209.12356.pdf>
<https://arxiv.org/pdf/2301.13848.pdf>

Few-Shot Learning: GPT-3



Fine-Tune LLMs

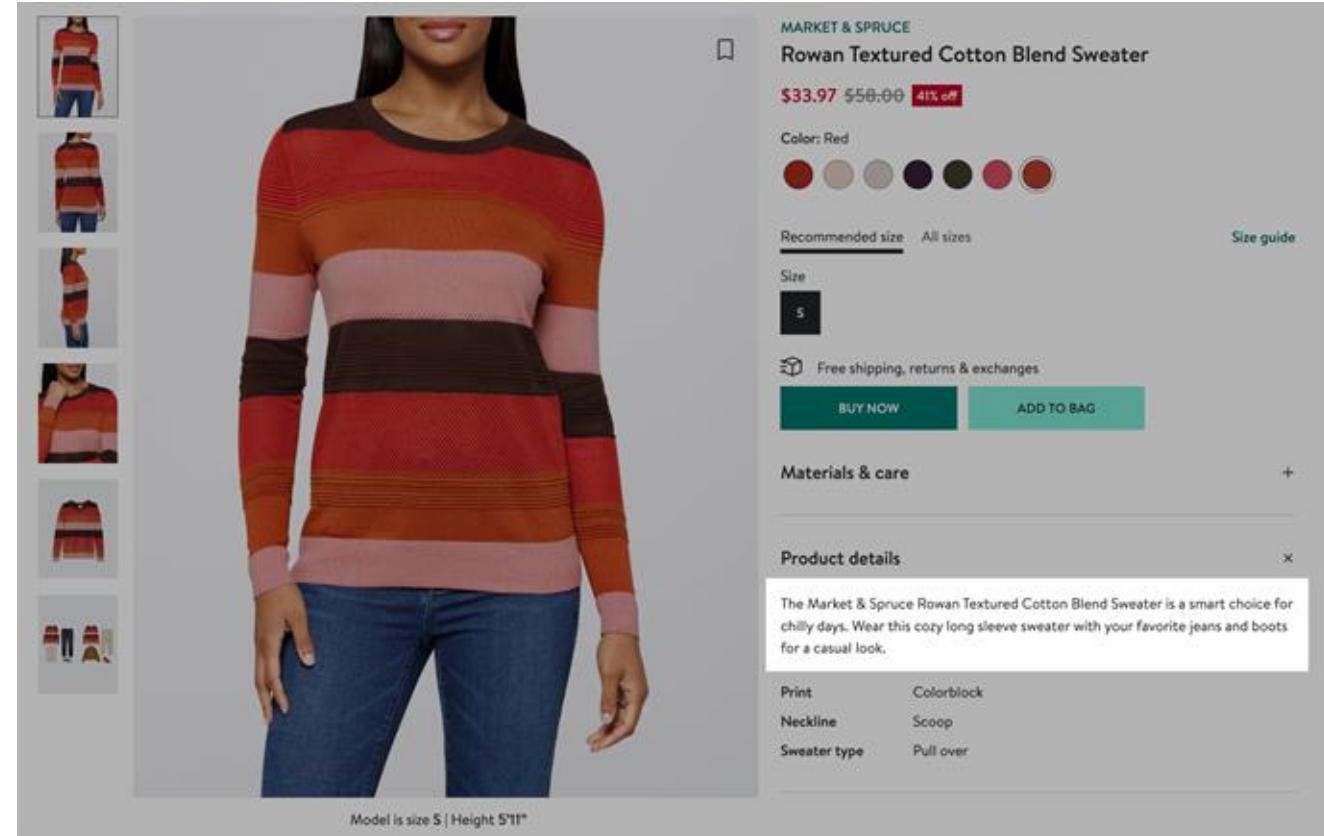
Train the model
with many
examples

```
{"prompt":"Overjoyed with the new iPhone! ->", "completion":" positive"}  
{"prompt":"@lakers disappoint for a third straight night https://t.co/38EFe43 ->", "completion":" negative"}  
{"prompt":"Happy with the new baby! ->", "completion":" positive"}  
{"prompt":"That movie was terrible last night| ->", "completion":" negative"}
```



Fine-Tuned LLMs

- Fine tuning with several hundred descriptions
- The resulting new descriptions were rated higher than human written examples



Current State of LLMs

Traditional niche domains of AI research:
semantic parsing, dialog, reasoning,
question answer, reading comprehension,
summarization

ARE DEAD!!



Armen Aghajanyan - Research Scientist at Meta AI

Rajiv Shah @rajistics

LLMs across common NLP tasks

LLMs: GPT-3,
FLAN-T5

Models:
BERT

Do we need traditional models?





Analytical

LLMs across common NLP tasks

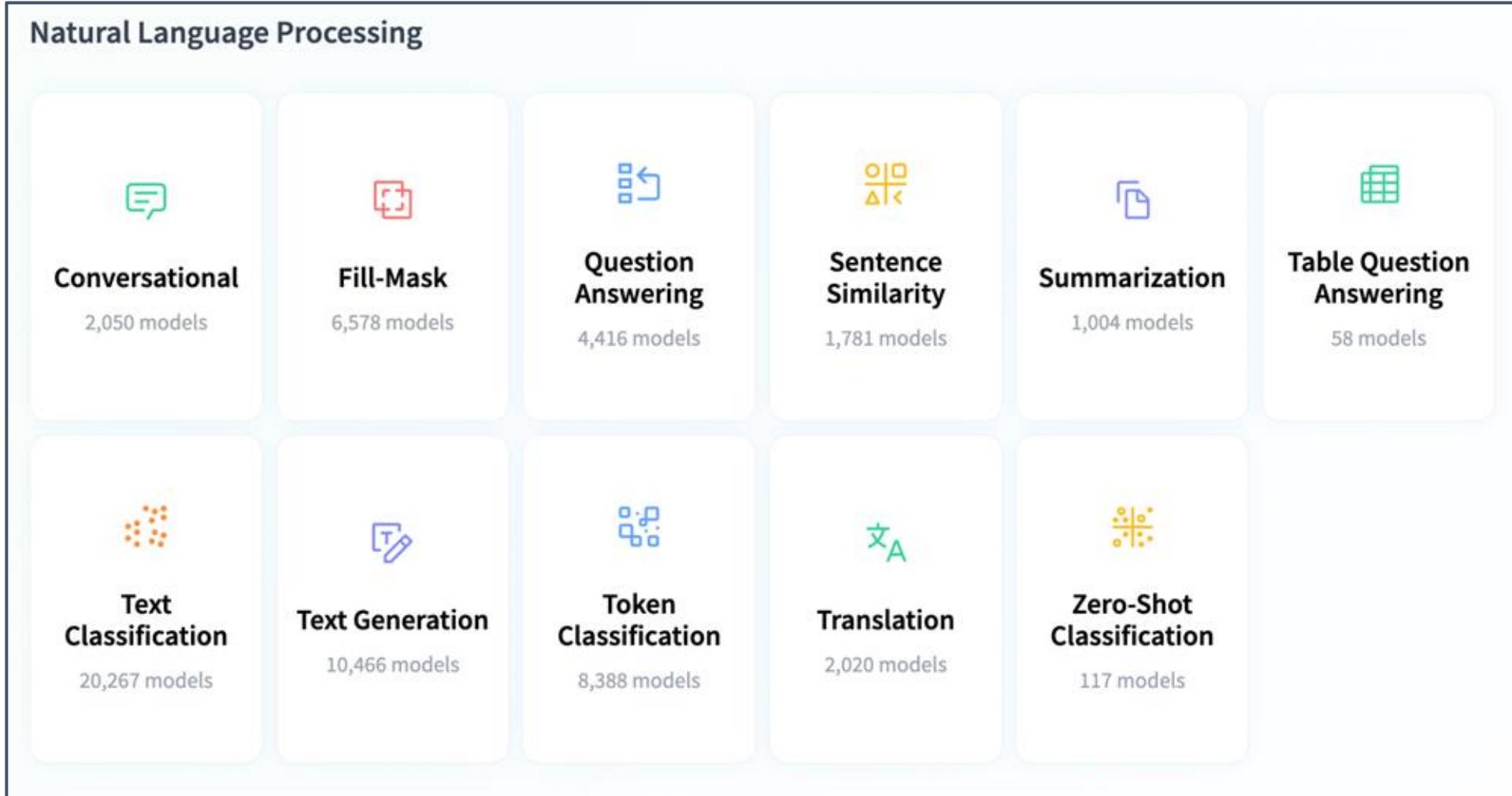
LLMs: GPT-3,
FLAN-T5

Models:
BERT

What are you trying to solve?



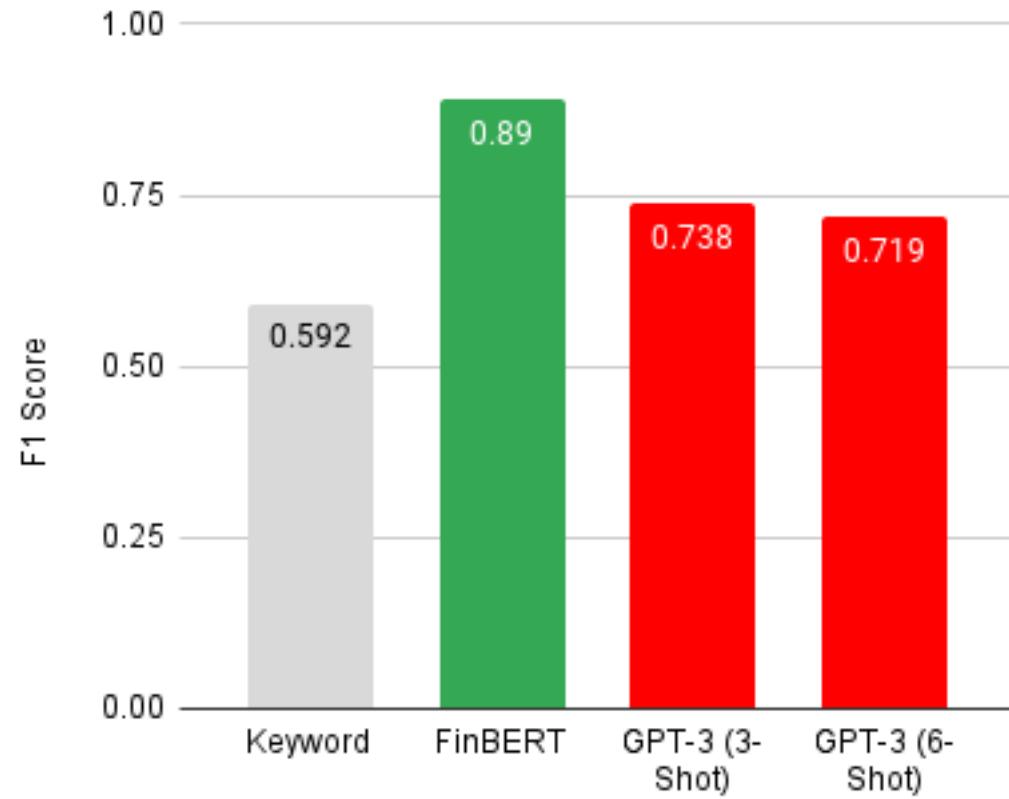
Traditional NLP Pretrain Models



Text Classification (Finance)

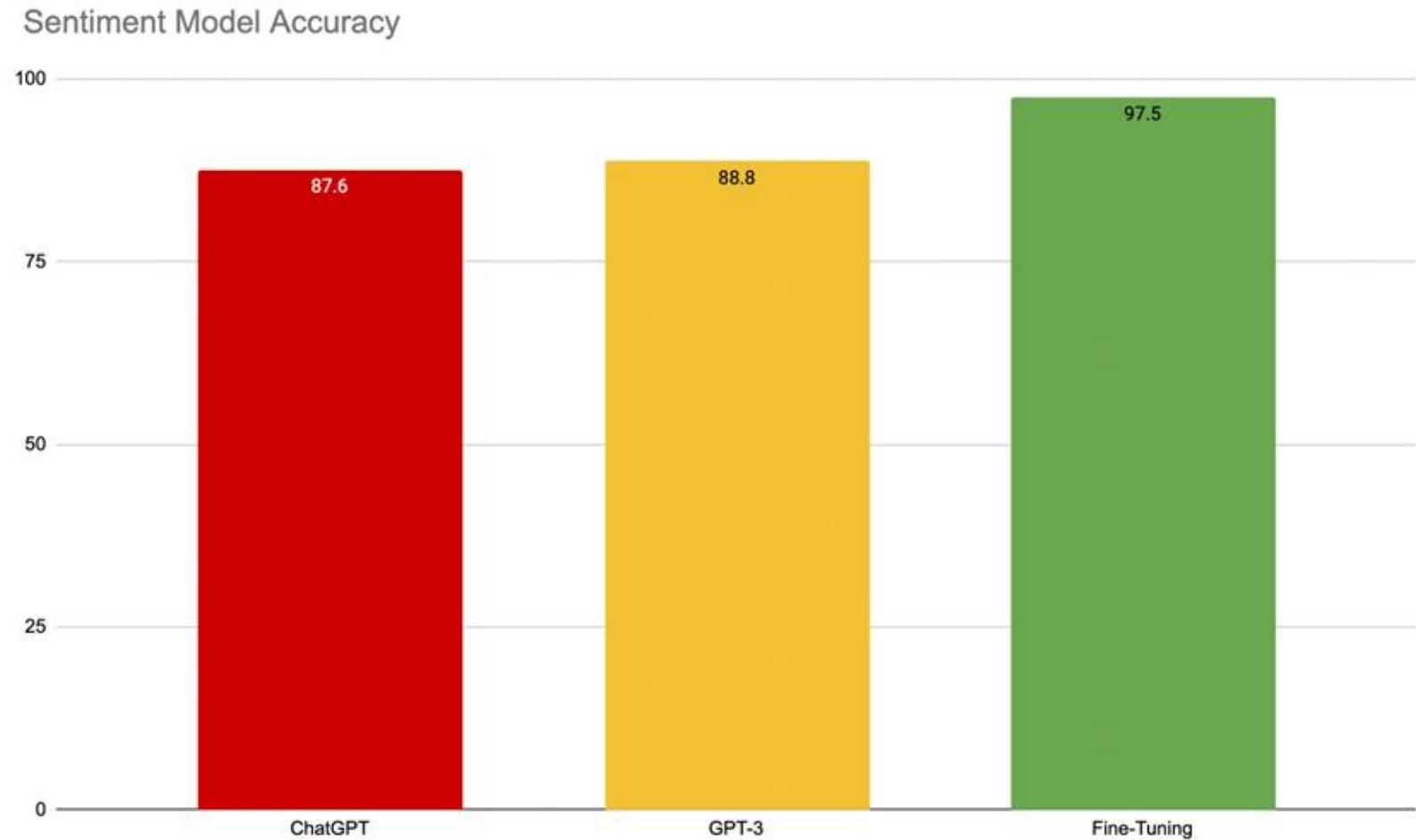
- Financial domain
- Over a 15% Drop in Accuracy

Financial Sentiment Analysis



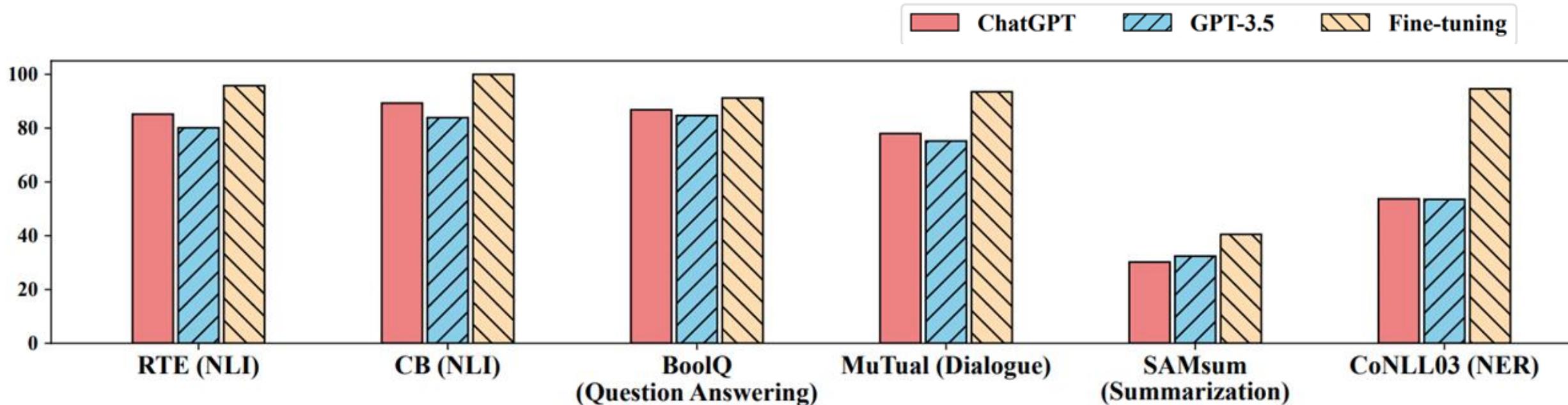
Text Classification (General)

- Fine Tuning
still better
than zero-
shot



LLMs across common NLP tasks

- Zero Shot versus Fine-Tuned Models



LLMs across common NLP tasks

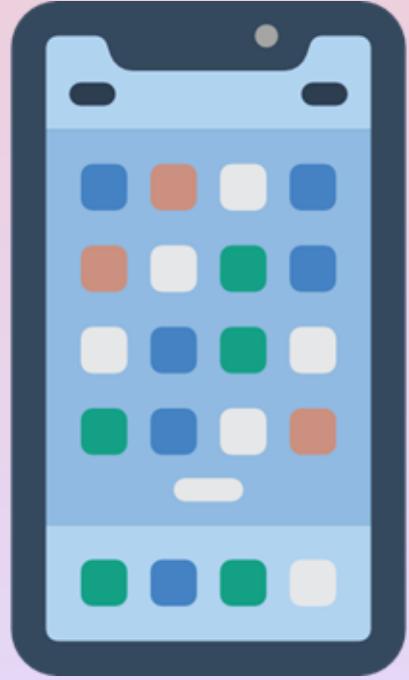
LLMs: GPT-3,
FLAN-T5

Models:
BERT

Starting to see the tradeoffs?



LLMs across common NLP tasks



Generalist



Specialist

Easy Decisions

For high value use cases, data science team will build a dedicated model

- For domain specific tasks -> Dedicated Model
- For high scale/low latency -> Dedicated model
- Explainability / Model Risk -> Dedicated model



Factors: LLMs vs. Model

- Predictive performance
- Scaling to large data
- Speed of Inference
- Data privacy
- Explainability
- Model risk for your organization
- Cost



Development and retraining time from your team

Operationalizing in your enterprise

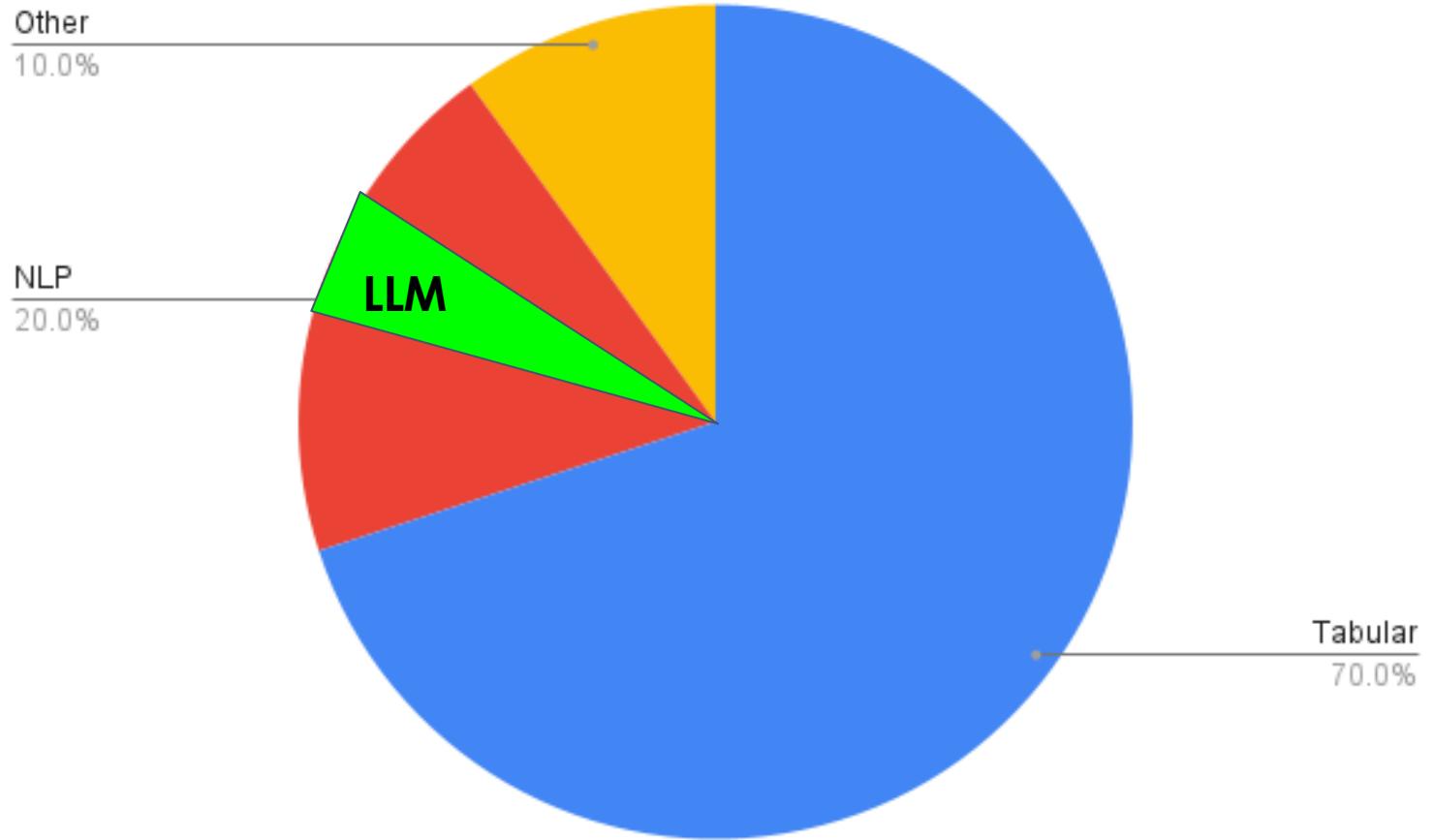
Rajesh Venkatasubramanian

Immediate Impact of LLMs

LLMs will likely have
low impact on the
current enterprise
analytics roadmap



Rajiv Shah @rajististics





ChatGPT Sprints to One Million Users

Time it took for selected online services to reach one million users



* one million backers ** one million nights booked *** one million downloads
 Source: Company announcements via Business Insider/LinkedIn



Rajiv Shah @rajis

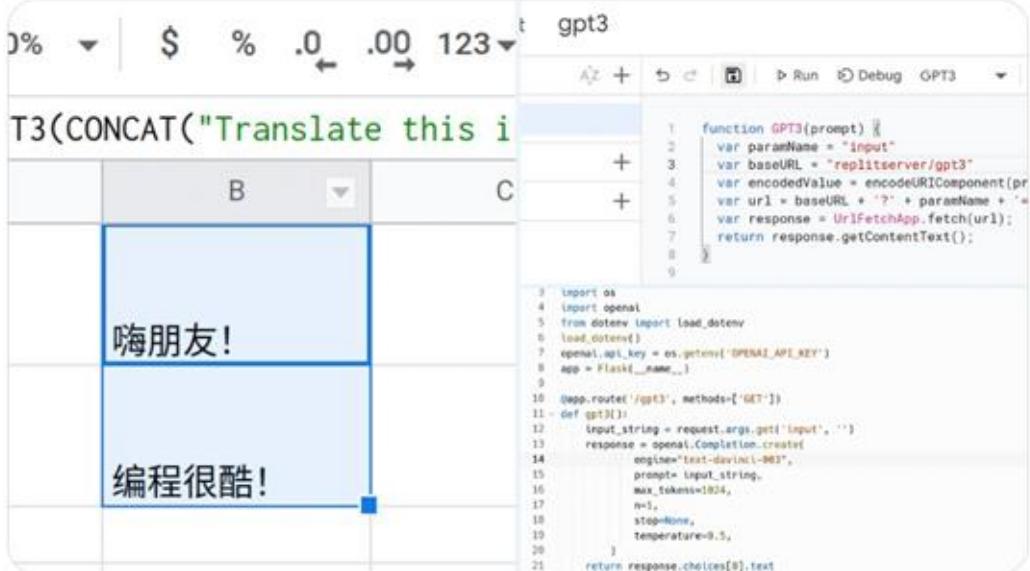


statista

Lot of Enthusiasm for using LLMs

 **John Horton** 
 @johnjhorton

With [@Replit](#), Google Sheets & [@OpenAI](#) API, you can hack together a spreadsheet-based, AI-powered workflow in like ~10 minutes. I'm going to teach my students this next week



```

function GPT3(prompt) {
  var params = "input"
  var baseURL = "replitserver/gpt3"
  var encodedValue = encodeURIComponent(prompt)
  var url = baseURL + "?" + params + "=" + encodedValue
  var response = UrlFetchApp.fetch(url)
  return response.getContentText()
}

import os
import openai
from dotenv import load_dotenv
load_dotenv()
openai.api_key = os.getenv('OPENAI_API_KEY')
app = Flask(__name__)

@app.route('/gpt3', methods=['GET'])
def gpt3():
  input_string = request.args.get('input', '')
  response = openai.Completion.create(
    engine="text-davinci-003",
    prompt=input_string,
    max_tokens=1024,
    n=1,
    stop=None,
    temperature=0.5,
  )
  return response.choices[0].text
  
```

7:08 PM · Mar 6, 2023 · 116.3K Views

64 Retweets 3 Quote Tweets 593 Likes



Rajiv Shah @rajistics

GPT3() Example Usage

File Edit View Insert Format Data Tools Extensions Help Last e...

Share

B2

A B C D E

	A	B	C	D	E
1	Messy Address Data		State Abbr.	Zip Code	
2	249 commonwealth Ave apt 9 boston Massachusetts 02116				
3	2731 Turtle Shores Dr, Bloomfield Hills, MI 48302				
4	133 5th Ave, Apt 2 Brooklyn, New York				
5	13902 High St, Whittier, CA 90605				
6	2323 Valley Street, # 307, Oakland California 94612				
7	2918 N Pine Grove Ave, Apt 3, Chicago, IL 60657				
8	415 4th St, Apt 3, Brooklyn, NY 11215				
9	1301 Aliceanna St, Apt 1403, Baltimore, Maryland 21231				
10	Apt. 3R, 724 Newkirk Avenue, Brooklyn, NY 11230				
11	3101 Wilson Street, San Francisco, California				
12	1512 Elm St, Los Angeles, CA 90012				
13	100 S 4th St, Philadelphia 19106				
14	867 Washington Ave, New York, NY 10011				
15	1660 Hollow Oak Drive, Apt 5, San Diego, CA 92105				
16	715 Washington St. New York, NY 10007				

701.3K views

0:00 / 2:16

Data Sanitization

Explore

Productivity Boost

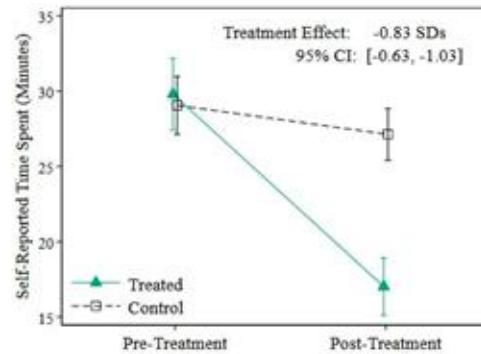
Raises productivity for mid-level professional writing tasks



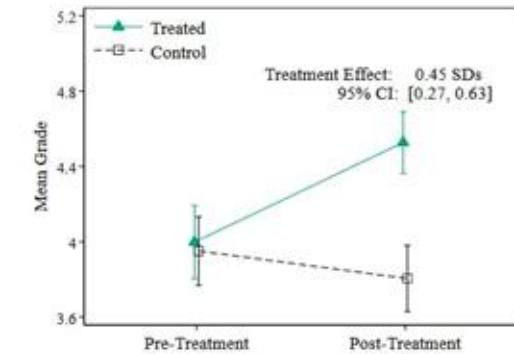
https://economics.mit.edu/sites/default/files/inline-files/Noy_Zhang_1.pdf

Rajiv Shah @rajististics

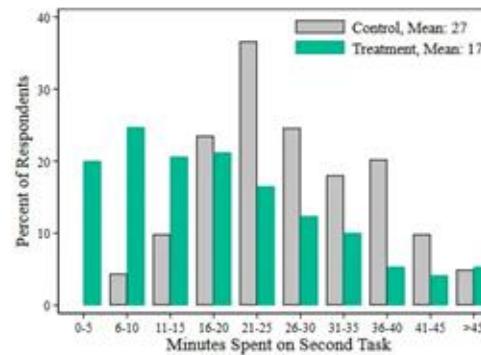
(a) Time Taken Decreases



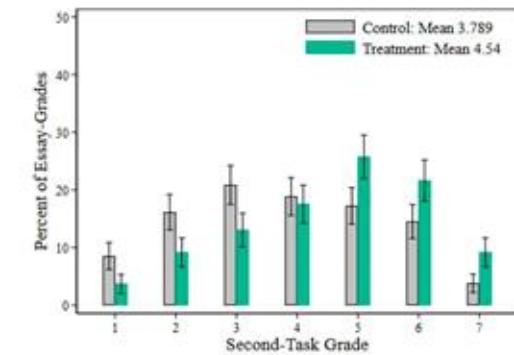
(b) Average Grades Increase



(c) Time Distribution (Second Task)

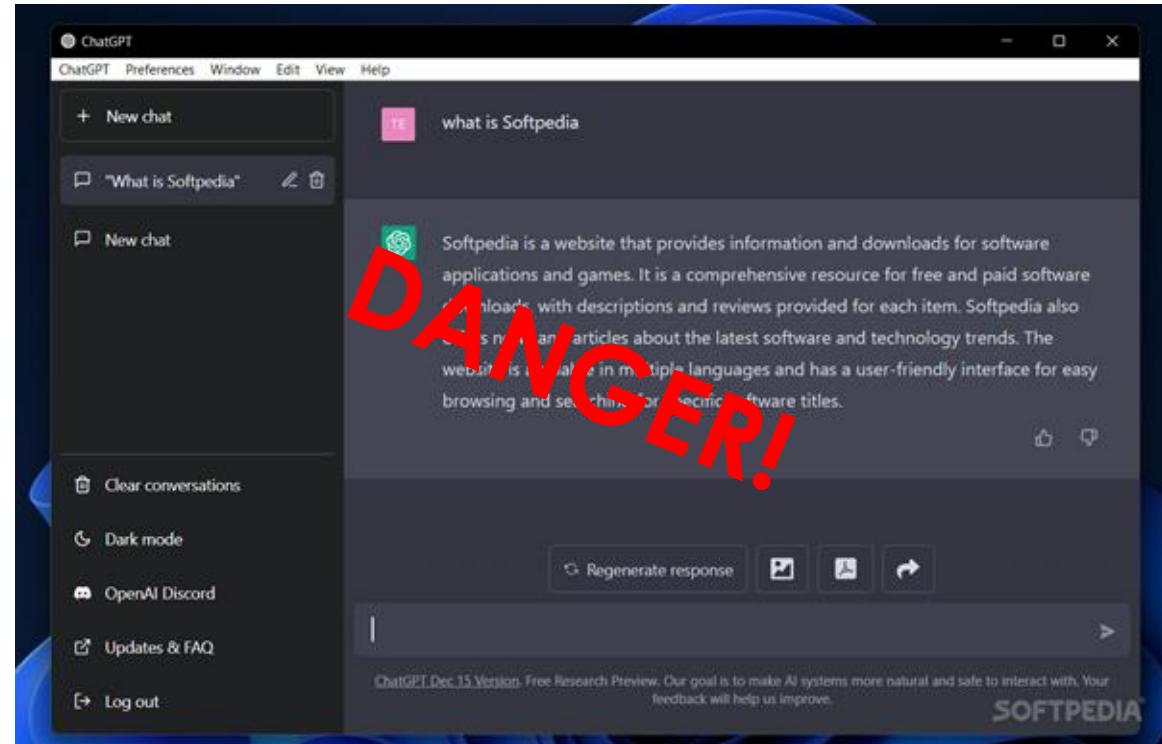


(d) Grades (Second Task)



People like to use LLMs for answers

- People use it like a search engine - they like the quick answers
- Reinigorating Search



Rajiv Shah @rajistics

Wolfram Demo

Spaces: JavaXpert/Chat-GPT-LangChain like 313 • Running

App Files and versions Community 20

Chat Settings Formality Translate to Emotions Max words

Conversational Agent using GPT-3.5 & LangChain

Chatbot Please summarize Moby-Dick for my upcoming book report

Moby-Dick is a classic novel by Herman Melville, published in 1851. It tells the story of Ishmael, a sailor who joins a whaling expedition led by the enigmatic Captain Ahab. The novel follows Ishmael and his crewmates as they search for the white whale Moby-Dick, a quest that leads to tragedy and destruction. Along the way, the novel explores themes of fate, revenge, and the power of nature.

Perfect, now give me a summary of hitchhikers guide to galaxy

The Hitchhiker's Guide to the Galaxy is a science fiction comedy series written by Douglas Adams. It follows the adventures of Arthur Dent, a hapless Englishman who is saved from the destruction of Earth by his alien friend Ford Prefect. Together, they travel through space and time, encountering a variety of strange and humorous characters and situations. The series is known for its witty dialogue, absurdist humor, and exploration of philosophical themes.

Great, now imagine a novel which is a cross between these two novels and write a summary for such novel.

A novel combining the themes of Moby-Dick and The Hitchhiker's Guide to the Galaxy could be an exciting and humorous adventure. The story could follow a hapless protagonist, similar to Arthur Dent, who joins a whaling expedition led by a mysterious captain, similar to Ahab. Along the way, the protagonist would encounter a variety of strange and humorous characters and situations, as well as explore themes of fate, revenge, and the power of nature. The novel could be a thrilling and humorous journey through space and time, with a unique blend of science fiction and comedy.

What's on your mind?

What's the answer to life, the universe, and everything?

Send



Rajiv Shah @rajistics

Wolfram Demo

- It determines whether it should use an external tool
- Then it acts using that tool

Chatbot

What is 2 to the 30th power?

Thought: Do I need to use a tool? Yes
Action: PAL-MATH Action
Input: What is 2 to the 30th power?
Observation: 1073741824
Thought: Do I need to use a tool? No
AI: The answer is 1073741824.
The answer is 1073741824.



Rajiv Shah @rajistics

Wolfram Demo

- It determines whether it should use an external tool
- Then it acts using that tool

How much did it rain in SF tod

Thought: Do I need to use a tool? Yes
 Action: Search Action
 Input: How much did it rain in SF today?
 Observation: California Nevada River Forecast Center - Your government source of hydrologic/weather data and forecasts for California, Nevada, and portions of southern Oregon. According to the California Nevada River Forecast Center, it rained 0.02 inches in San Francisco today.
 Thought: Do I need to use a tool? No
 AI: According to the California Nevada River Forecast Center, it rained 0.02 inches in San Francisco today.
 According to the California Nevada River Forecast Center, it rained 0.02 inches in San Francisco today.



Rajiv Shah @rajistics

VIDEO LINK FOR TALK: <https://youtu.be/dKBD-3hnjW0>

Emerging LLM Capabilities

- Retrieval plugin - save info
- Browsing
- Code Interpreter
- Additional Plugins

Post Date Fact-check +

The publication date of the "Learning from human preferences" blog post is given at the top of the page. I will quote this excerpt to fact check the second date provided by the User.

Read content

Finished browsing

The OpenAI blog post on "Unsupervised Sentiment Neuron" was actually published on April 6, 2017¹, and the "Learning from human preferences" post was published on June 13, 2017². Therefore, approximately two months and one week passed between the two publications.

0 0 0

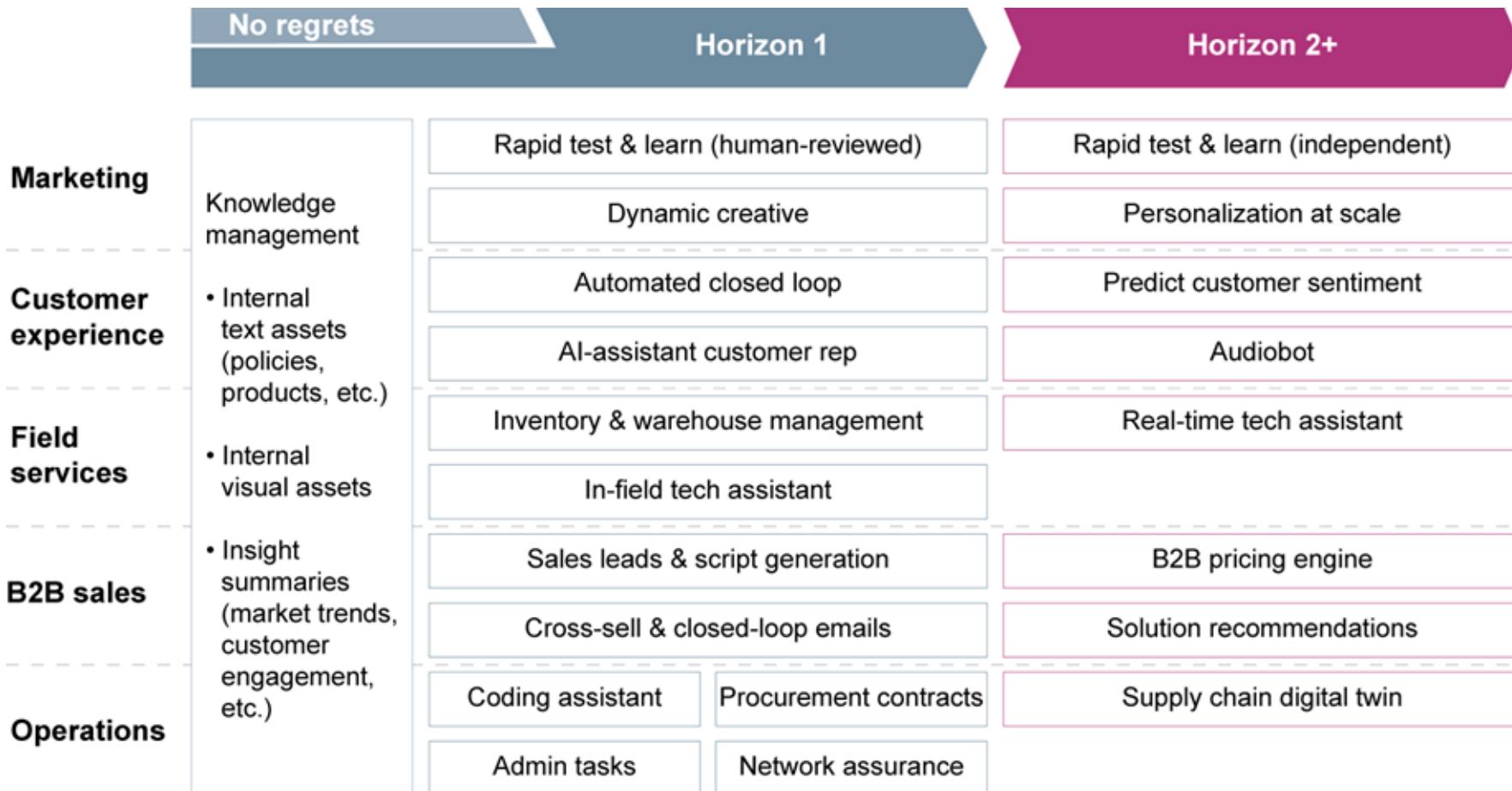


LLM Trends

- Emerging Capabilities
 - Code based LLMs
 - Multimodal LLMs (GPT-4 & MiniGPT-4)
 - Updated LLMs
 - Plugins / Services LLMs
 - Managing Security/Risk
 - LLMOps



Next Generation of AI Use Cases



Source: Bain & Company



Rajiv Shah @rajistics

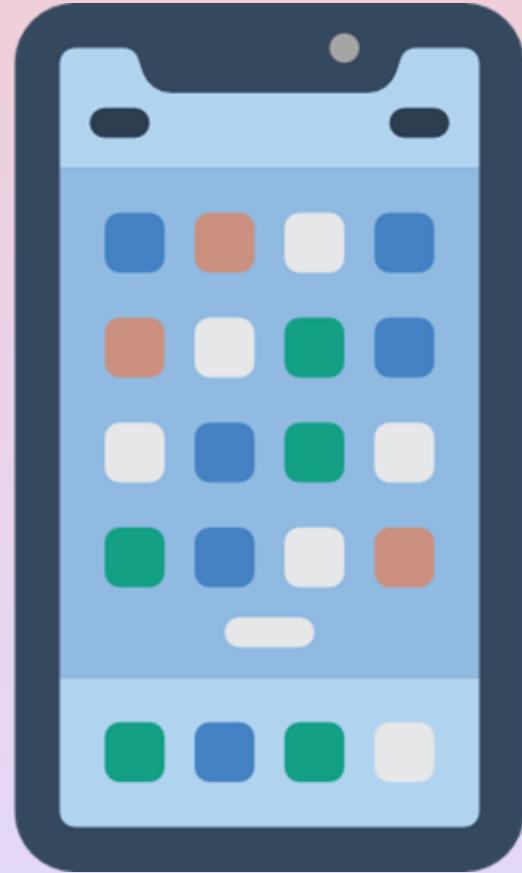
Transformational Impact



Rajiv Shah @rajistics







How do Large Language Models Do?

Playground **Compare**

who is Nat Friedman?



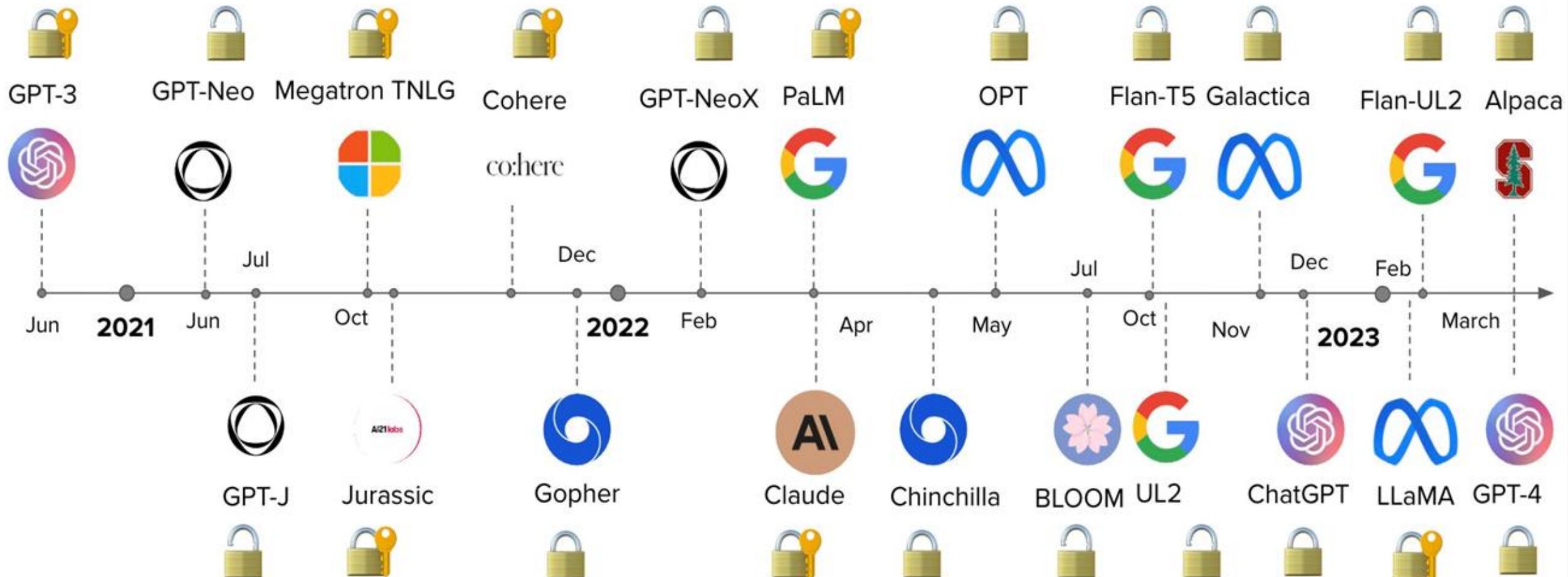
Submit

anthropic:claude	huggingface:google/flan-t5-xxl	huggingface:google/flan-ul2	cohere:command-medium-nightly	cohere:command-xlarge-nightly
Nat Friedman is the CEO of GitHub, a popular code hosting platform. Before becoming CEO of GitHub in 2018, Nat Friedman was a software engineer and entrepreneur. He is the creator of the Gnome desktop environment for Linux and co-founded Ximian, a company that built software for Linux and open source desktops. Friedman has been involved in open source software and Linux for over 20 years. As CEO of GitHub, he has focused on growing GitHub's business and expanding its use among developers.	nat friedman	National Security Advisor	Nat Friedman is a technology investor and entrepreneur who has focused on artificial intelligence and virtual reality. He is the co-founder of several companies, including Andreessen Horowitz, reached valued at nearly 80 billion dollars in 2017.	Nat Friedman is an American-Israeli entrepreneur and technologist who is best known as the co-founder and CEO of Xamarin, a mobile app development platform that was acquired by Microsoft in 2016 for \$400 million. Friedman is also a venture capitalist and angel investor, and has invested in companies such as Wix, GitHub, and WeWork. He previously worked at Microsoft as a developer on the Windows Server team and as a program manager on the Visual Studio team.
00:04	123 chars/s	494 chars	00:01	77 chars/s
494 chars	13 chars/s	13 chars	245 chars/s	462 chars
openai:gpt-3.5-turbo	openai:text-ada-001	openai:text-curie-001	openai:text-davinci-002	openai:text-davinci-003
Nat Friedman is the CEO of GitHub, a web-based hosting service for version control using git. He previously co-founded Xamarin, a platform for developing mobile apps, which was acquired by Microsoft in 2016.	Nat Friedman is an American businessman and entrepreneur. He is the co-founder and CEO of Comcast NBCUniversal, as well as a number of other businesses.	Nat Friedman is an American entrepreneur and investor. He is the co-founder and CEO of Recode, a news and technology company based in Silicon Valley.	Nat Friedman is the CEO of GitHub.	Nat Friedman is the CEO of GitHub, a software development platform. He was previously the CEO of Xamarin, a mobile app development platform, and prior to that he was the Vice President of Developer Platform at Microsoft.
00:02	103 chars/s	207 chars	00:00	111 chars/s
207 chars	154 chars/s	154 chars	36 chars/s	222 chars
00:00	151 chars/s	151 chars	00:02	222 chars



nat.dev

LLMs since GPT3



LLMs - Data Sources

9

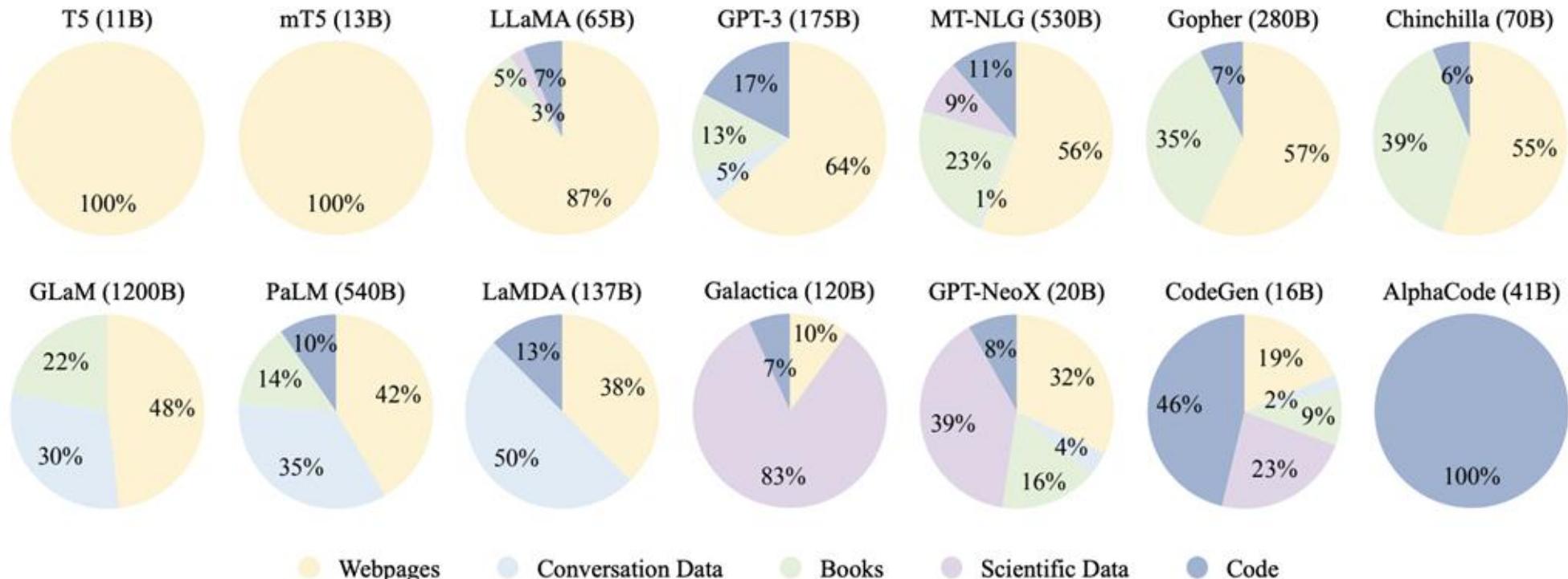


Fig. 2. Ratios of various data sources in the pre-training data for existing LLMs.



LLMs - Last Updated

TABLE 2
Statistics of commonly-used data sources.

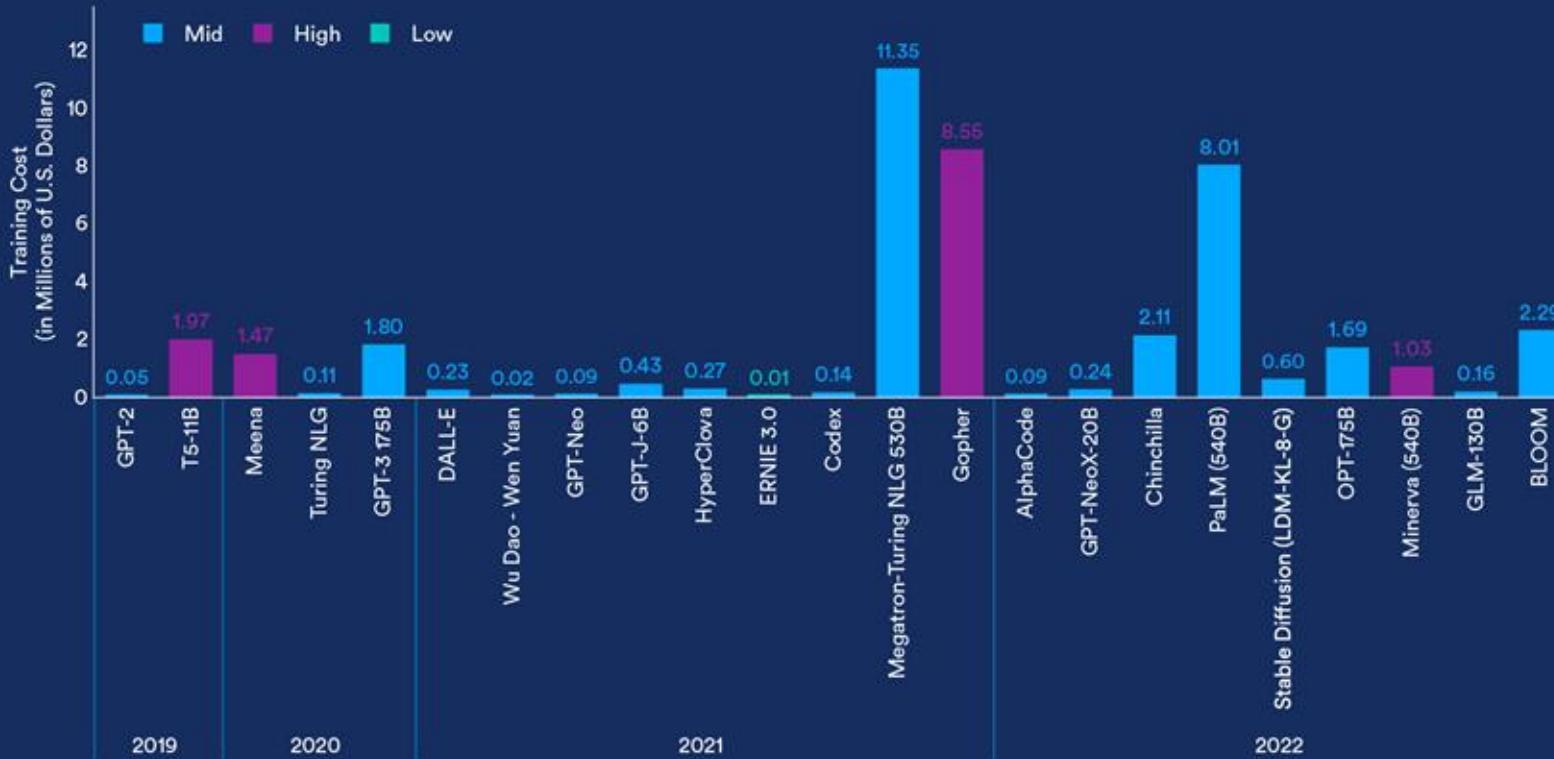
Corpora	Size	Source	Latest Update Time
BookCorpus [100]	5GB	Books	Dec-2015
Gutenberg [101]	-	Books	Dec-2021
C4 [71]	800GB	CommonCrawl	Apr-2019
CC-stories-R [102]	31GB	CommonCrawl	Sep-2019
CC-NEWS [27]	78GB	CommonCrawl	Feb-2019
REALNEWS [103]	120GB	CommonCrawl	Apr-2019
OpenWebText [104]	38GB	Reddit links	Mar-2023
Pushift.io [105]	-	Reddit links	Mar-2023
Wikipedia [106]	-	Wikipedia	Mar-2023
BigQuery [107]	-	Codes	Mar-2023
the Pile [108]	800GB	Other	Dec-2020
ROOTS [109]	1.6TB	Other	Jun-2022



LLM Training Cost



Estimated Training Costs of Large Models



Source: AI Index, 2022 | Chart: 2023 AI Index Report

LLMs

Train your
Own LLM

Open
Source

Commercial
LLMs

API



BloombergGPT

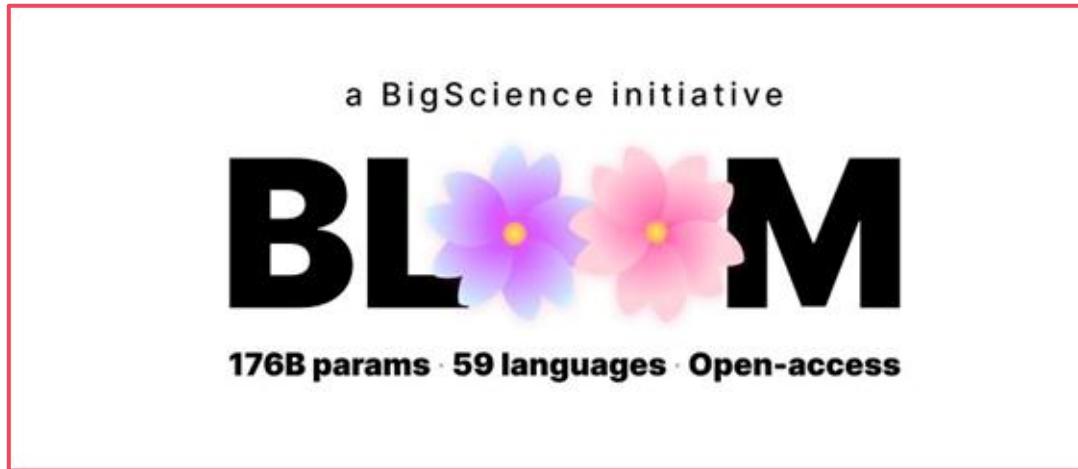
- 50B Parameter Model
 - 345B tokens from general purpose datasets
 - 363B token dataset based on Bloomberg's data sources
 - Trained on 64×8 A100 40GB
- 

	BLOOMBERGGPT	GPT-NeoX	OPT _{6B}	BLOOM _{176B}
ConvFinQA	43.41	30.06	27.88	36.31
FiQA SA	5.07	10.59	51.60	53.12
FPB	31.07	41.64	48.67	50.25
Headline	82.20	73.22	79.11	76.51
NER	60.12	60.90	57.49	55.56
All Tasks (avg)	62.51	51.90	53.01	54.35
All Tasks (WPE)	0.93	0.27	0.33	0.47

6 months
Sophisticated team
\$300k in compute

Table 3: Results on financial domain tasks.

Open Source LLMs



- Open source Flamingo (not yet publicly available)

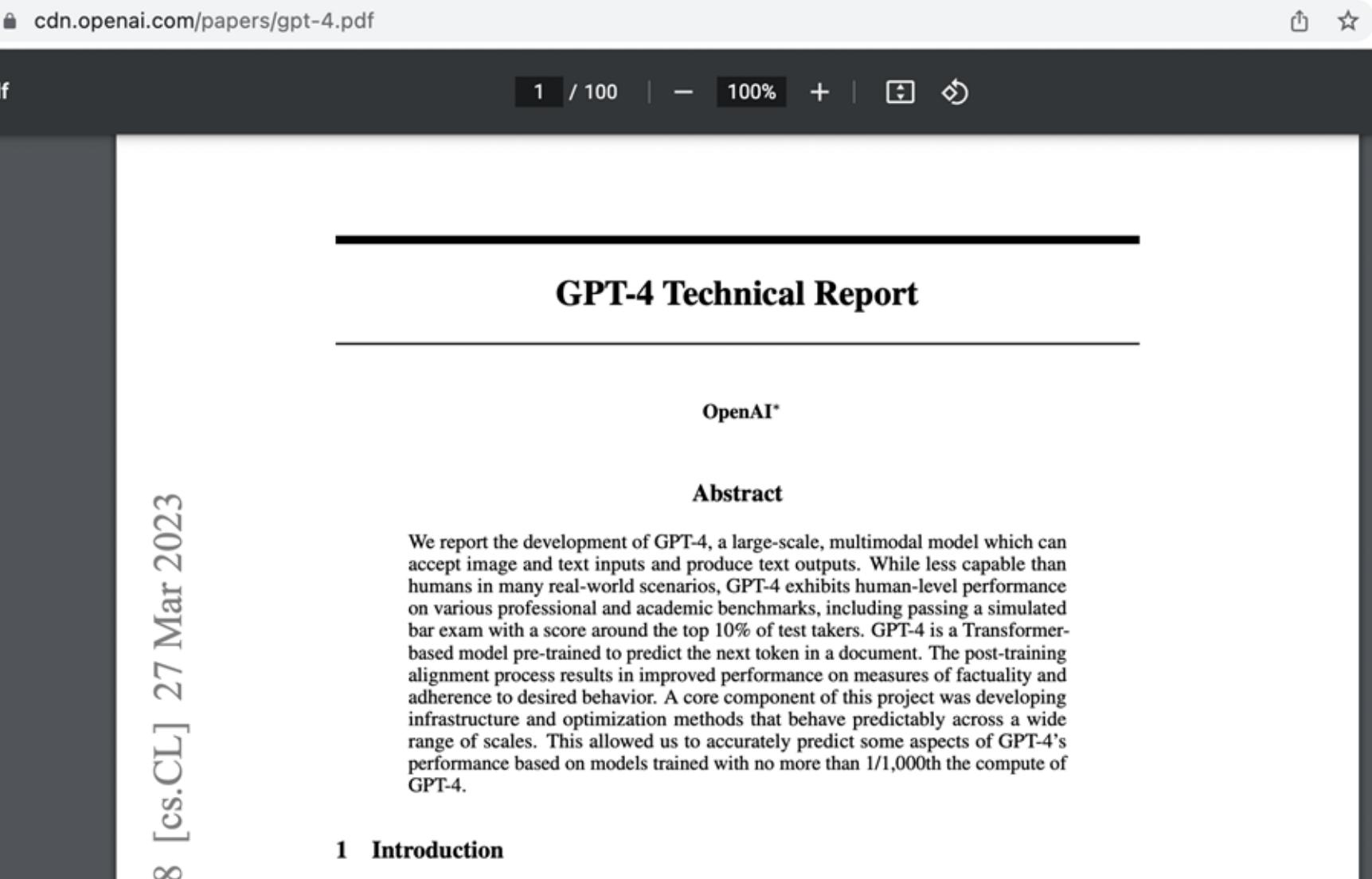


Open Source LLMs



- [LLaMA](#)
 - o Models ranging from 7B to 65B
 - o Commercial usage not allowed
 - [Alpaca](#)
 - o LLaMa fine-tuned on the “Alpaca dataset”, a dataset of 52k examples of instruction tuning generated by GPT-3.
 - o However, the Alpaca dataset contains many inefficiencies which are being fixed in [this repository](#). The cleaned version is [available on the HF hub](#).
 - o Commercial usage not allowed
 - Flan UL2
 - o Apache 2.0 license
 - Flan T5
 - o Apache 2.0 license
 - GPT Neo X/Pythia
 - o Open-Assistant fine-tuned Pythia (12B parameters) on 22k instructions:
<https://huggingface.co/OpenAssistant/oasst-sft-1-pythia-12b>.
 - GPT-J
 - o Databricks fine-tuned GPT-J on the Alpaca dataset and open-sourced the model, which they called [Dolly](#).
 - BLOOM/BLOOMZ
 - which are being fixed in [this repository](#). The cleaned version is
- Outdated**

Commercial LLMs



The screenshot shows a PDF document titled "GPT-4 Technical Report" by OpenAI. The document is a single page with a dark header bar containing navigation controls. The main content includes the title, author information, an abstract, and the first section of the report. The URL "cdn.openai.com/papers/gpt-4.pdf" is visible in the browser's address bar.

cdn.openai.com/papers/gpt-4.pdf

1 / 100 | - 100% + | ↻ ↺

GPT-4 Technical Report

OpenAI*

Abstract

We report the development of GPT-4, a large-scale, multimodal model which can accept image and text inputs and produce text outputs. While less capable than humans in many real-world scenarios, GPT-4 exhibits human-level performance on various professional and academic benchmarks, including passing a simulated bar exam with a score around the top 10% of test takers. GPT-4 is a Transformer-based model pre-trained to predict the next token in a document. The post-training alignment process results in improved performance on measures of factuality and adherence to desired behavior. A core component of this project was developing infrastructure and optimization methods that behave predictably across a wide range of scales. This allowed us to accurately predict some aspects of GPT-4's performance based on models trained with no more than 1/1,000th the compute of GPT-4.

1 Introduction



Evaluation

“We’re talking about ChatGPT and we know nothing about it,” she said. “It could be three raccoons in a trench coat.”



Source: Sasha Luccioni @SashaMTL

?????

This report focuses on the capabilities, limitations, and safety properties of GPT-4. GPT-4 is a Transformer-style model [39] pre-trained to predict the next token in a document, using both publicly available data (such as internet data) and data licensed from third-party providers. The model was then fine-tuned using Reinforcement Learning from Human Feedback (RLHF) [40]. Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar.



VIDEO LINK FOR TALK: <https://youtu.be/dKBD-3hnjW0>

Benchmarks for LLMs

- Stanford HELM Report -
 - Many of these automated test suites rely on multiple choice answer prompts, as open question free-form questionnaires are difficult to evaluate.
- GPT-4 Technical Report included human tests (GRE)



<https://crfm.stanford.edu/helm/latest/>
<https://arxiv.org/abs/2303.08774>

AI21 labs

BigScience

Google

Microsoft



ANTHROPIC

co:here



Meta

NVIDIA

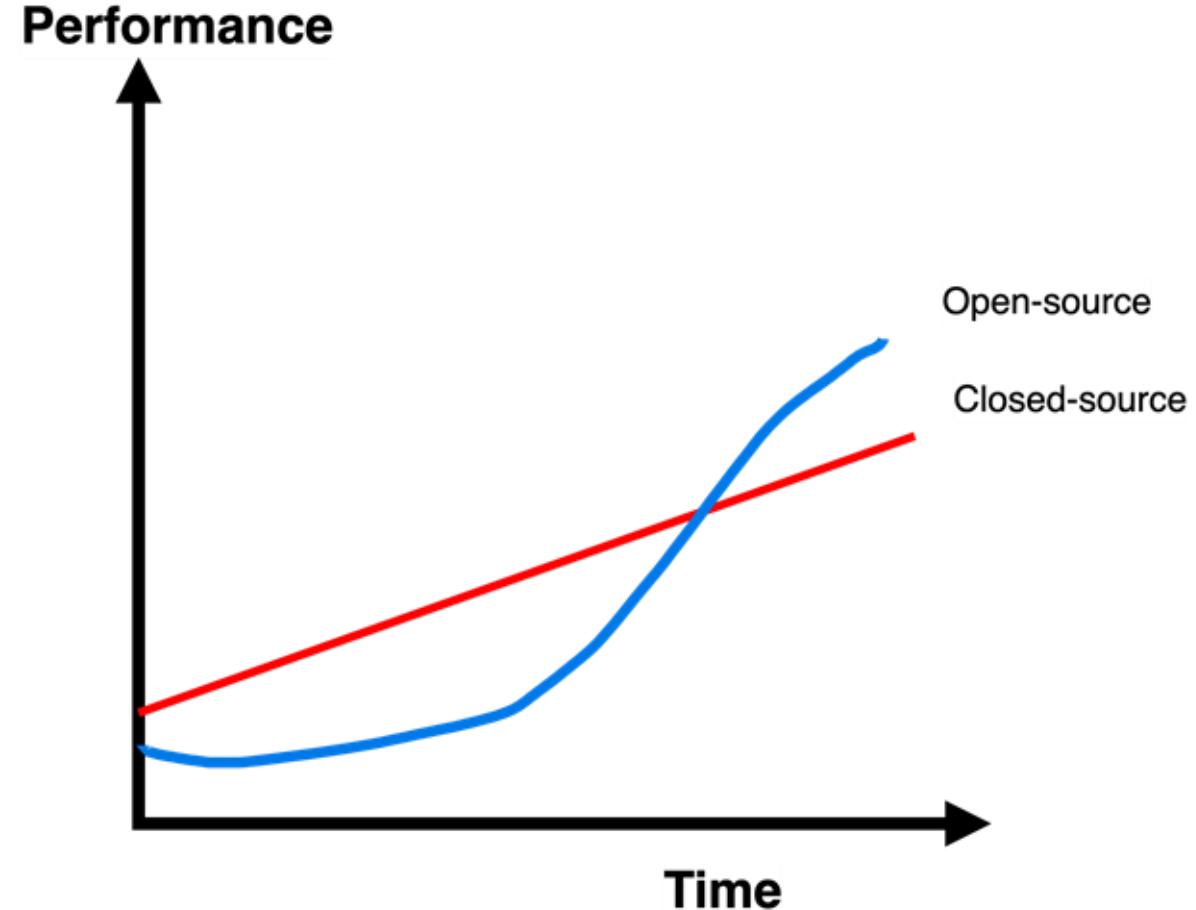
OpenAI

Yandex

TOGETHER

Historical Understanding

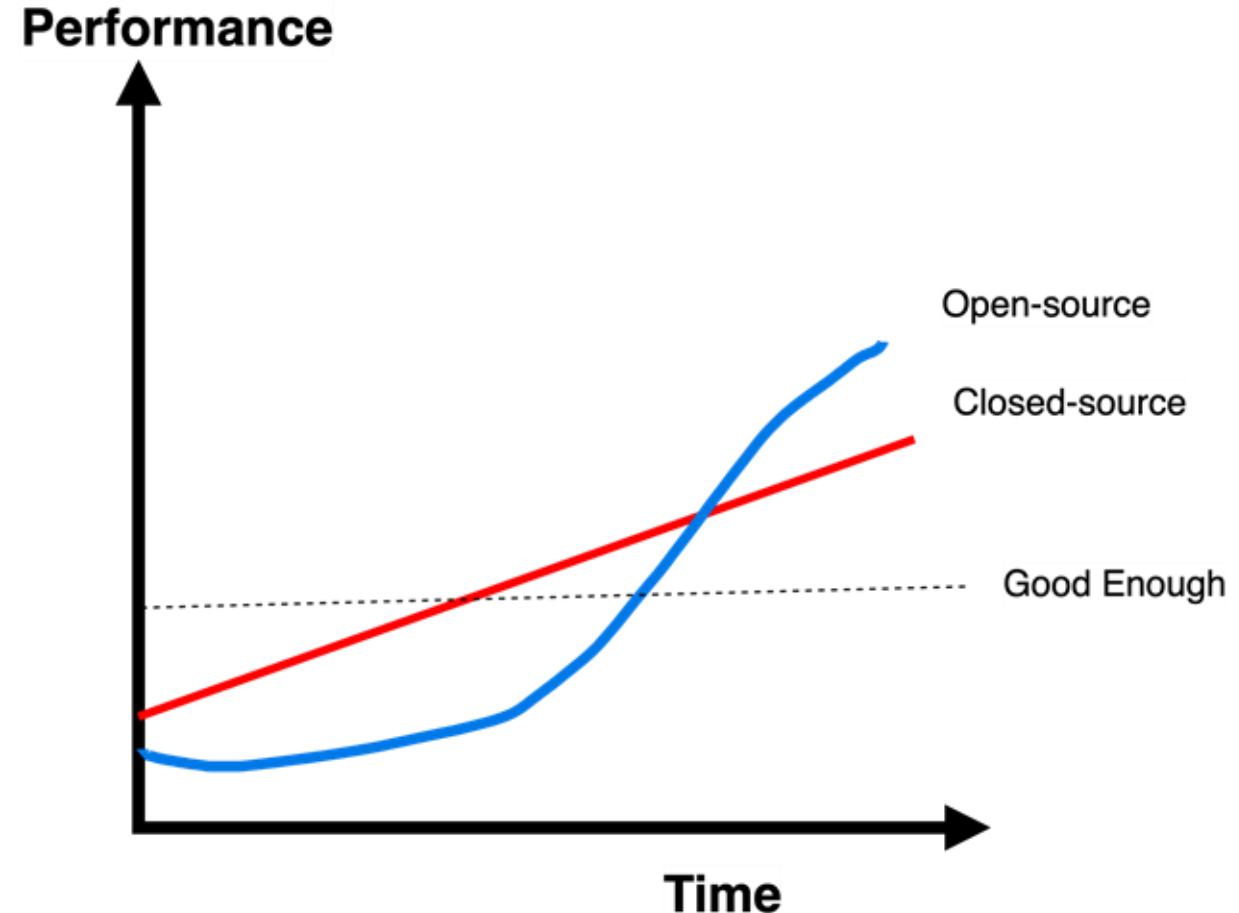
- Open source is showing momentum with new releases / updates weekly
- The gap is closing with recent releases of OpenAssistant and Vicuna



Source: Abubakar Abid

Good Enough

- At some point we will hit diminishing returns with the performance of LLMs
- Proprietary models will need to keep pushing that up



LLMs Considerations

- Predictive Performance
- Last Updated
- Training Dataset (e.g., Code or Financial News)
- Size of LLM
- Cost
- Inference speed
- Run locally or is this an API
- Ownership/control of the model
- Model Risk / Governance / Ethical Concerns



Time to Market

Reference: [Sharon Bress - Best Practices for Training LLMs from Scratch - Weights & Biases](https://sharonbress.github.io/best-practices-for-training-llms-from-scratch/)



Vicki

@vboykis

...

Something I think about often wrt LLMs is what [@NeuralBricolage](#) said in her wonderful NormConf talk, “You need to own the means of production.” If you don’t own the model and don’t own the API you are building on sand.



Sharon Zhou  @realSharonZhou · 13h

Overheard at an AI event...

audience: how do I build anything defensible on LLMs?

speakers: *crickets*

audience: *laughs*

The actual answer is just to ~build your own LLM~.

Buuuut this chart show what that's like right now. Don't forget your error handler ;)

What'll be that... [Show more](#)

Build Your Own LLMs 🦄					
Model	Keep Your Data	Permissive License	Easy-To-Use Library	Reliable* API	*90.000%+
?	Yes	Yes	Yes	Yes	Yes
GPT-4	No	No	Yes	Yes	Yes
LLaMA	Yes	No	No	No	No
Pythia	Yes	Yes	No	No	No
Bard	No	No	No	No	No

7:19 AM · Apr 12, 2023 · 4,966 Views

Multi-LLM Landscape

- Enterprises will probably need to several different diverse LLMs for their tasks
 - General purpose LLM
 - Domain specific LLMs
 - Fast inference LLMs





Analytical



Transformative

When should we use LLMs?



Which LLM?



Analytical

When should
we use LLMs?

Start simple: internal chatbot or
search app using an LLM

Which LLM?

Start simple: Start with a vendor
API and an open source model
with (prompting and fine tuning)

Transformative

Many new generative use
cases (unknowable now)

Multiple LLMs from open
source, pretrained, and
commercial LLMs





**Applied
Tools for LLMS**



**Fundamental
R&D for LLMs**

A Practical Perspective on LLMs for Enterprises



Rajiv Shah



@rajistics

raj@huggingface.co

VIDEO LINK FOR TALK: <https://youtu.be/dKBD-3hnjW0>

