

# Enterprise & LLMs

March 2023

Rajiv Shah | @rajistics  
raj@hf.co



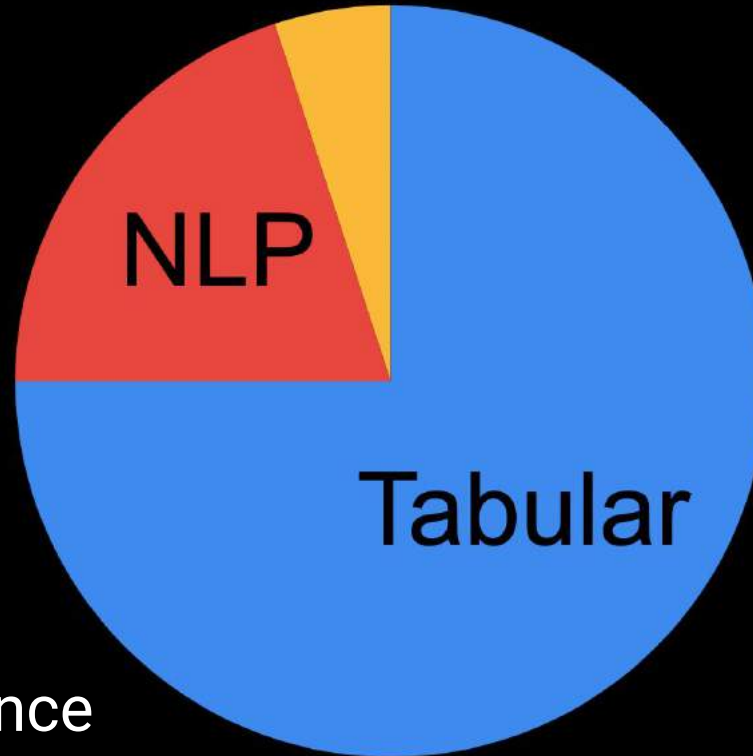
# Overview

- NLP in the Enterprise
- Effect of LLMs
  - Existing NLP use case
  - New opportunities
- What you should do today



# Machine Learning Use Cases in the Enterprise

- Tabular
- NLP
- Other



Source: Raj's Experience

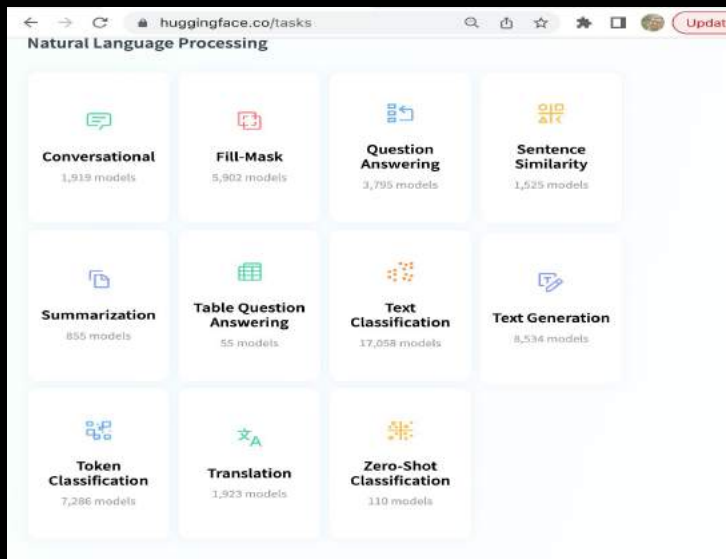


# Will LLMs Displace Traditional NLP Models?



# Ranking Interest in NLP Use Cases

## Tasks

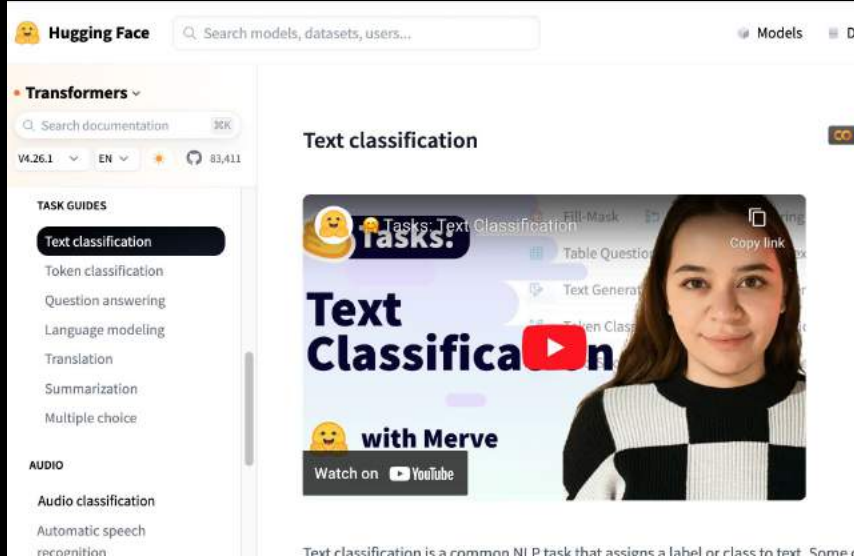


1. Text Generation
2. Question Answering
3. Text Classification
4. Summarization
5. Sentence Similarity



# Ranking Interest in NLP Use Cases

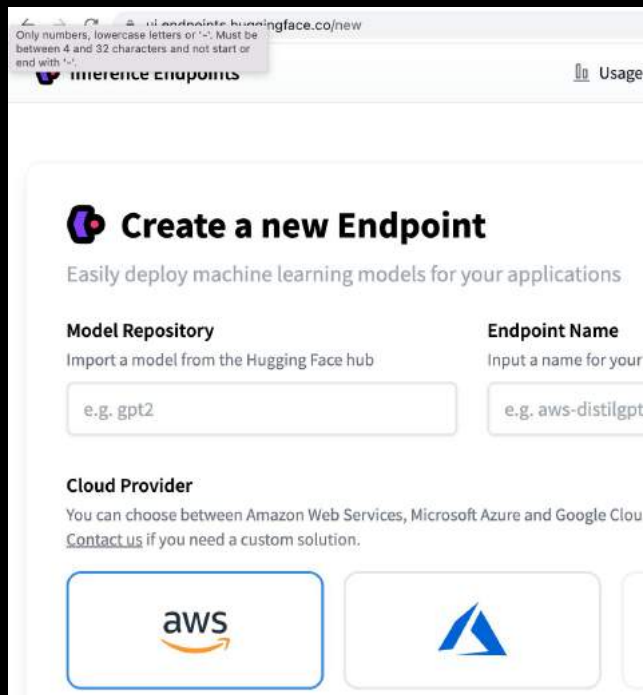
Docs



1. Text Classification
2. Summarization
3. Question Answering
4. Token Classification



# Ranking NLP deployments



The screenshot shows the 'Create a new Endpoint' form on the Hugging Face Inference Endpoints page. The form includes a header with the Hugging Face logo and the title 'Create a new Endpoint'. Below the title is a subtitle 'Easily deploy machine learning models for your applications'. The form is divided into three main sections: 'Model Repository' with a text input field containing 'e.g. gpt2', 'Endpoint Name' with a text input field containing 'e.g. aws-distilgpt', and 'Cloud Provider' with three radio button options: 'aws' (selected), 'Microsoft Azure', and 'Google Cloud'. A 'Contact us' link is visible below the cloud provider options.

1. Text Classification
2. Sentence Embeddings
3. Text Generation
4. Token Classification
5. Question Answering
6. Summarization

Source: Hugging Face Deployment Stats



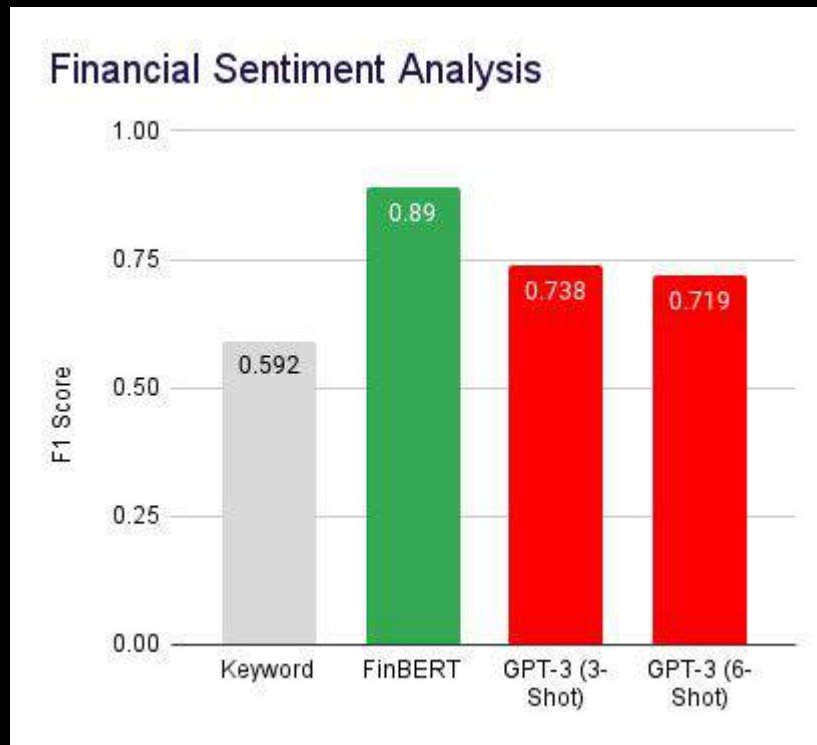
# Let's Compare LLMs to Traditional NLP Models





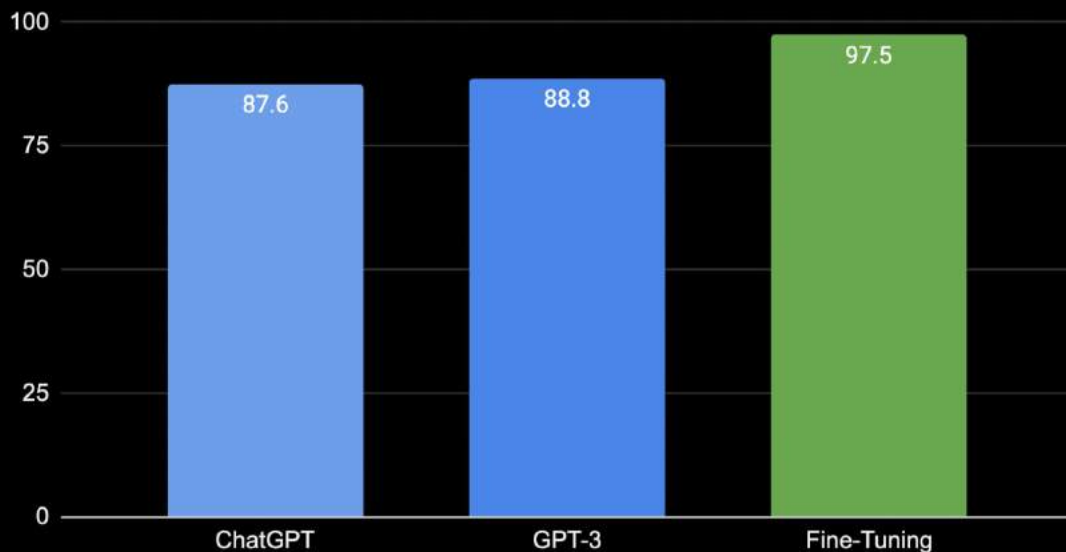
# LLMs for Text Classification

- Financial domain
- Over a 15% Drop in Accuracy



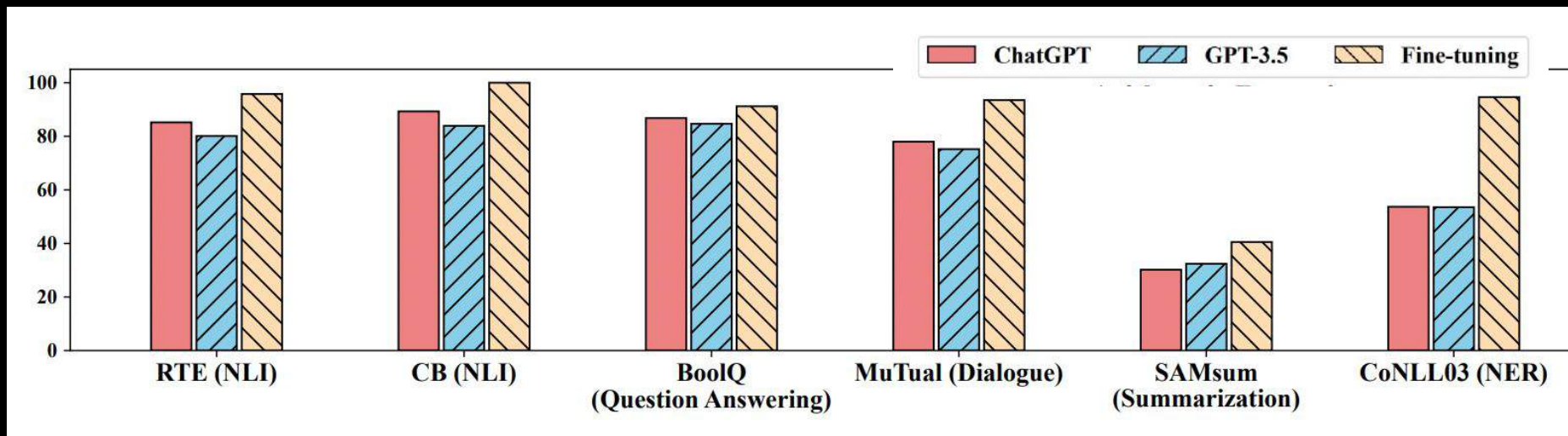
# LLMs for Text Classification

Sentiment Model Accuracy



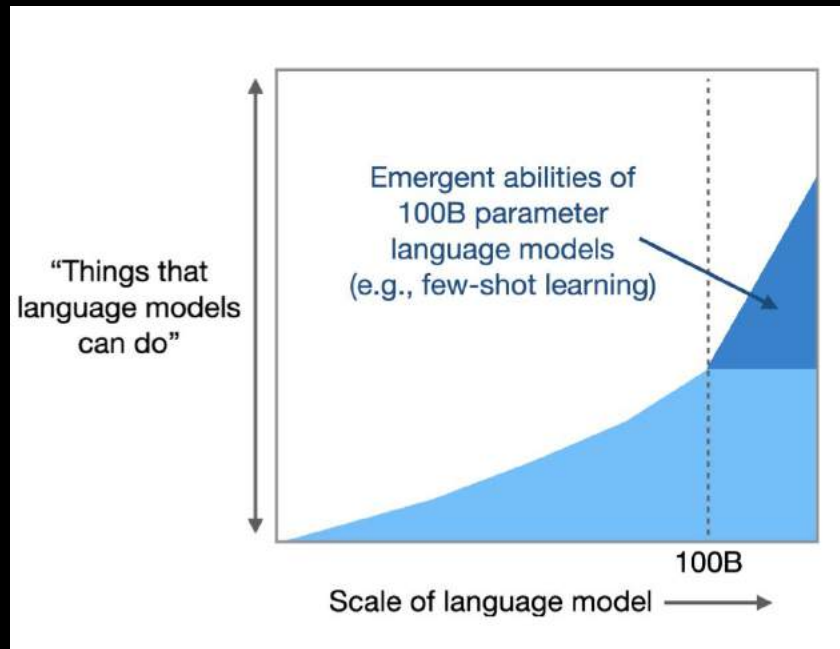
# LLMs across common NLP tasks

- Zero Shot versus Fine-Tuned Models



# Few Shot Learning

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



# Few Shot Learning

- Give the model a few examples
- Then use it to predict in the same way

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

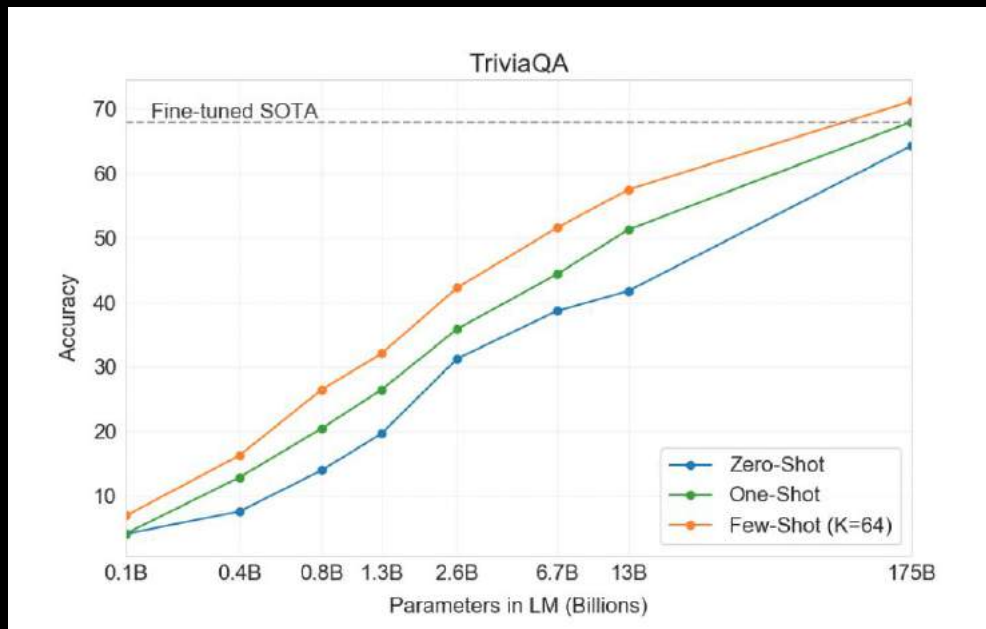
Article: Murtagh a stickler for success Northeastern field hockey coach Cheryl Murtagh doesn't want the glare of the spotlight that shines on her to detract from a team that has been the America East champion for the past three years and has been to the NCAA tournament 13 times.

Label:



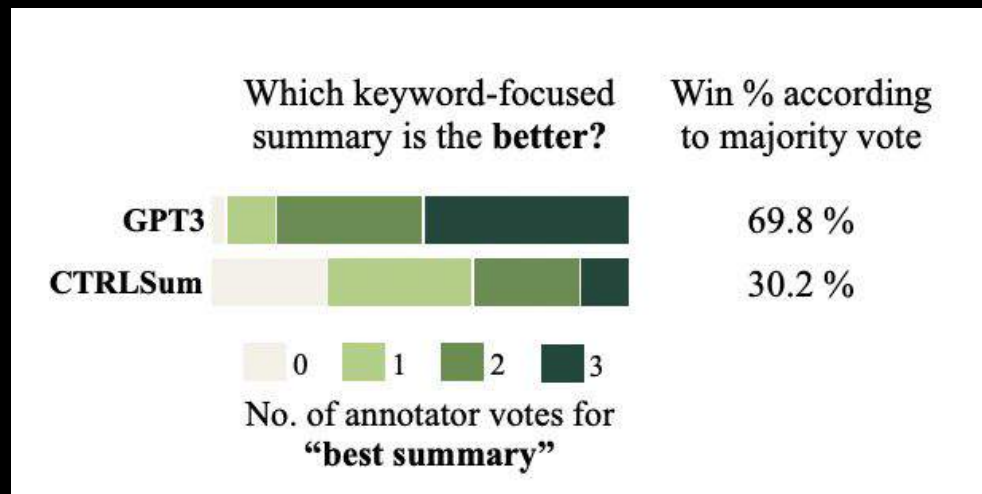
# Few-Shot Learning - GPT-3

- GPT-3:  
nearly matching  
the performance  
of state-of-the-art  
fine-tuned systems



# Few-Shot Learning - News

News Summarization:  
GPT-3 is better than  
fine-tuning & close to  
human level



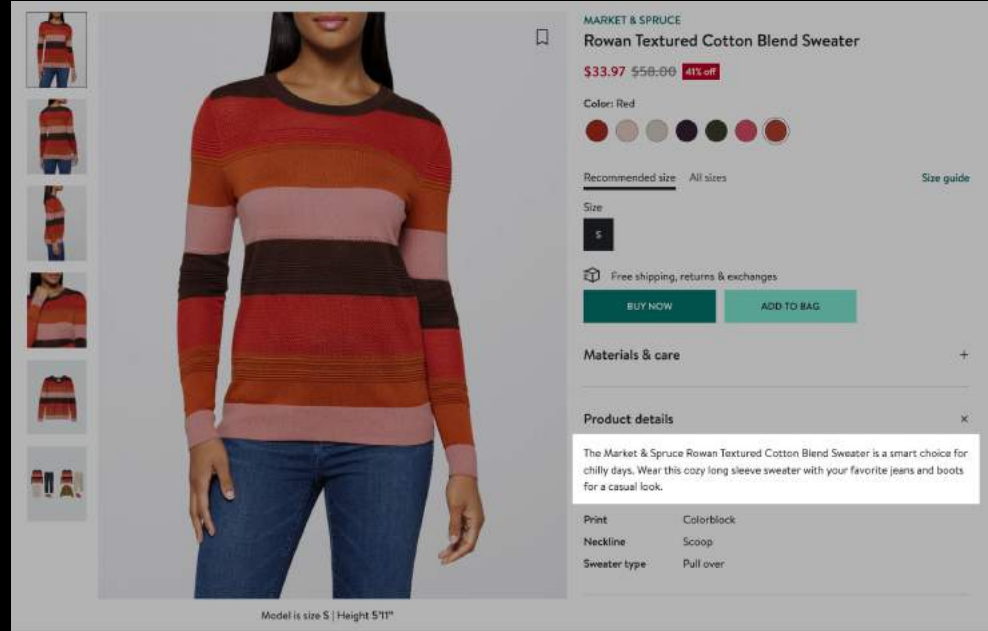
<https://arxiv.org/pdf/2209.12356.pdf>  
<https://arxiv.org/pdf/2301.13848.pdf>



# Fine Tuned LLMs

## Product Descriptions:

AI Generated descriptions from several hundred fine tuned examples were rated higher than human written examples



<https://multithreaded.stitchfix.com/blog/2023/03/06/expert-in-the-loop-generative-ai-at-stitch-fix/>





# How do Large Language Models Do?



# When to use LLMs?

For domain specific tasks -> Fine Tuning wins

For high scale/sensitive applications -> Dedicated open source model

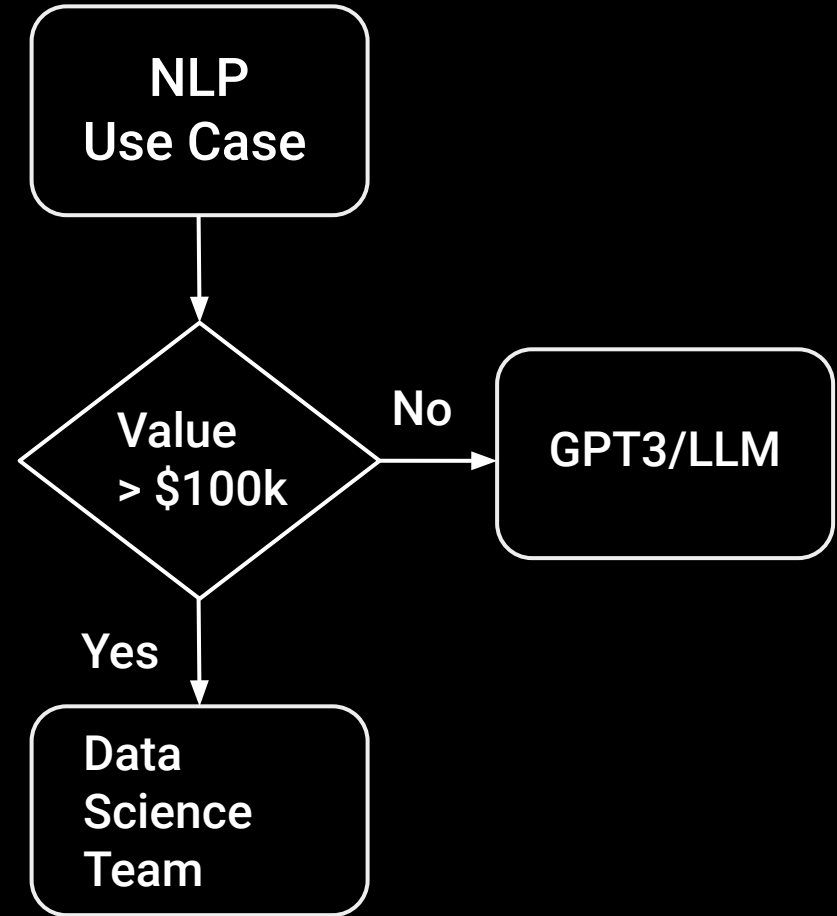
For many general tasks:

- Few shot performance is close
- Zero shot performance is still good
- This gap is likely to close quickly



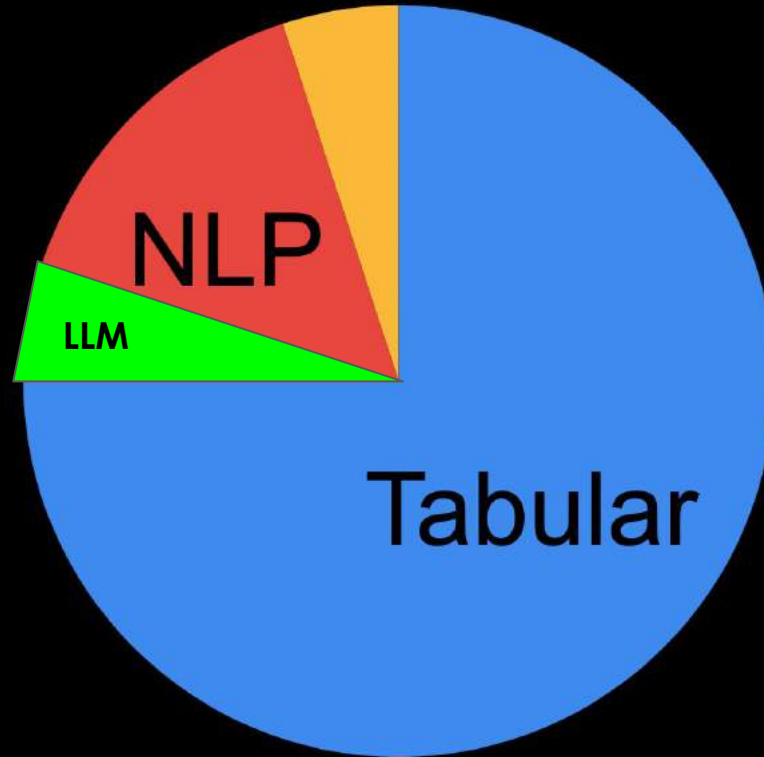
# Impact of LLMs on Existing NLP Use Cases

*Encourage using LLMs by educated users for the low value use cases.*



# Impact of LLMs on Existing NLP Use Cases

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





# User Perspective

# 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



# Lot of Enthusiasm for using LLMs



John Horton ✓

@johnhorton

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

The screenshot shows a Replit workspace with two main panels. On the left is a Google Sheet titled 'gpt3'. The visible part of the sheet has a header row with 'T3(CONCAT("Translate this i' and two columns, B and C. Cell B2 contains the text '嗨朋友!' (Hi friend!) and cell C2 contains '编程很酷!' (Coding is so cool!). On the right is a code editor with a JavaScript file named 'gpt3.js'. The code defines a function 'GPT3(prompt)' that uses the OpenAI API to translate the input prompt. It also includes a REST client setup for a Replit server, a route handler for the '/gpt3' endpoint, and a GET method that calls the 'GPT3' function and returns the result.

```
function GPT3(prompt) {  
  var paramName = 'input'  
  var baseUrl = 'replitserver/gpt3'  
  var encodedValue = encodeURIComponent(prompt)  
  var url = baseUrl + '?' + paramName + '=' + encodedValue  
  var response = UrFetchApp.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
```

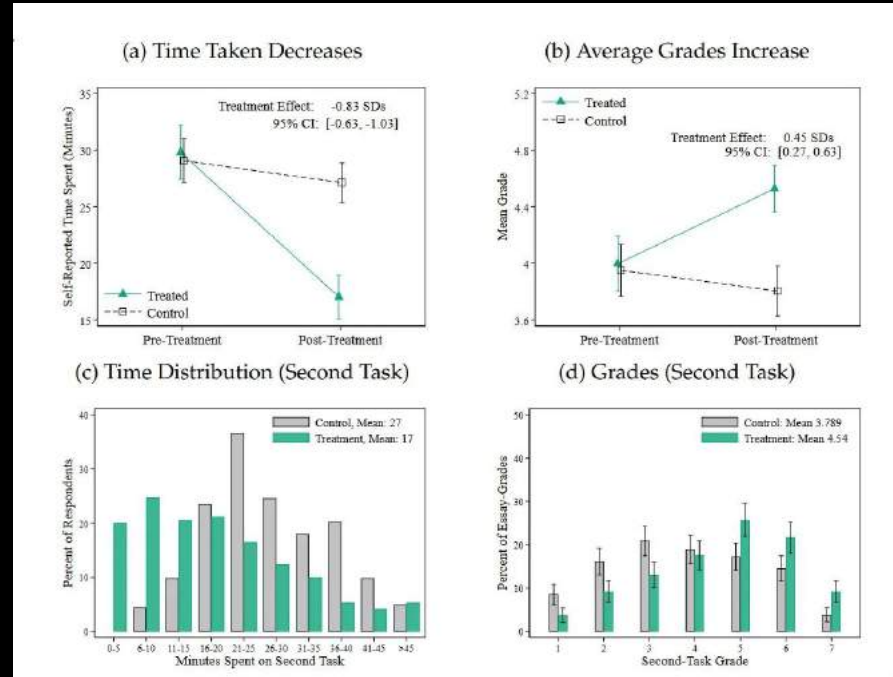
7:08 PM · Mar 6, 2023 · 116.3K Views

64 Retweets 3 Quote Tweets 593 Likes



# Using LLMs

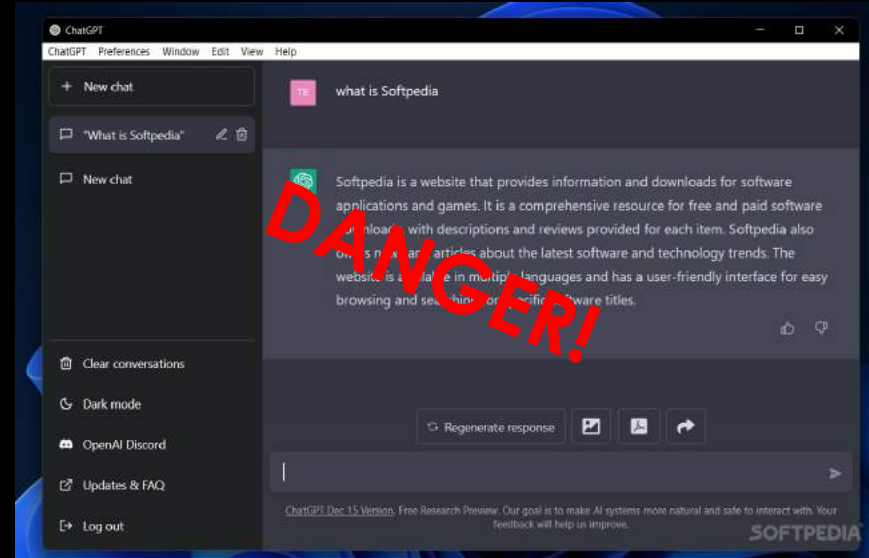
Raises productivity for  
mid-level professional  
writing tasks







# People like to use LLMs for answers

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



# Proper Way to use LLMs for Search

- Information Retrieval + LLM
-  LlamaIndex  (GPT Index)
  - Central interface to connect your LLM's with external data.



# LLMs as Decision Makers

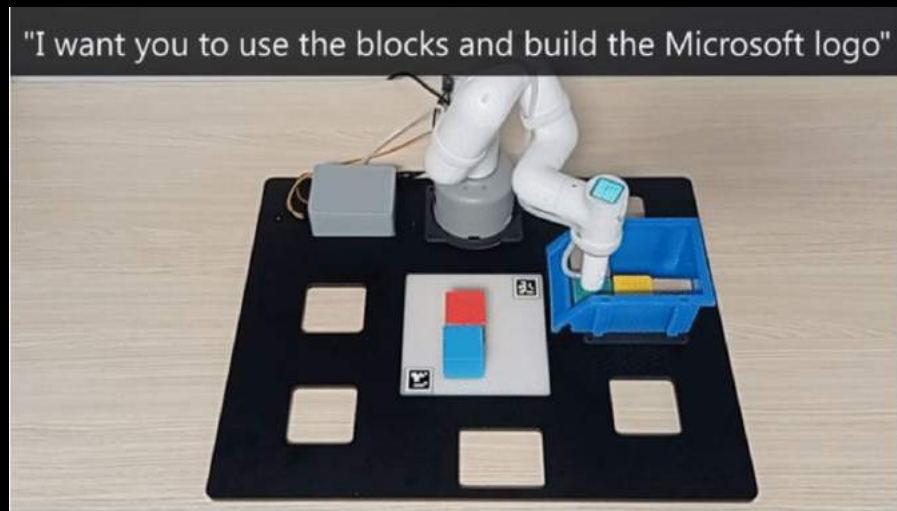
- LLMs can make intermediate decisions (Agent)
  - Intent classification
- LLMs can generate text to connect to APIs and services

```
> Entering new AgentExecutor chain...  
  I need to find out who Leo DiCaprio's girlfriend  
  is and then calculate her age raised to the 0.43  
  power.  
Action: Search  
Action Input: "Who is Leo DiCaprio's girlfriend?"  
Observation: Camila Morrone  
Thought: I need to find out Camila Morrone's age  
Action: Search  
Action Input: "How old is Camila Morrone?"  
Observation: 25 years  
Thought: I need to calculate 25 raised to the  
0.43 power  
Action: Calculator  
Action Input: 25^0.43  
  
> Entering new LLMMathChain chain...  
25^0.43  
```python  
import math  
print(math.pow(25, 0.43))  
```
```



# LLMs Can Work with Many Diverse Services

- Extending ChatGPT to control robots



<https://www.microsoft.com/en-us/research/group/autonomous-systems-group-robotics/articles/chatgpt-for-robotics/>



# Recommendations for Enterprise Analytics

- **Start building knowledge of using LLMs**
  - Building prompting skills / LangChain
  - Start educating the rest of the organization on how to use LLMs properly
    - Identify the crucial use cases
    - Start collecting data on the interactions your users have
  - Prepare for multiple LLM landscape (OpenAI, Cohere, Hugging Face, . . . )



# Recommendations for Enterprise Analytics

- **Use this excitement to prioritize search related projects.**
  - Start building Information Retrieval systems
  - Have that prompt engineer integrating multiple systems
- **Start looking at other services/API that can interconnect with LLMs**



# Enterprise & LLMs

March 2023

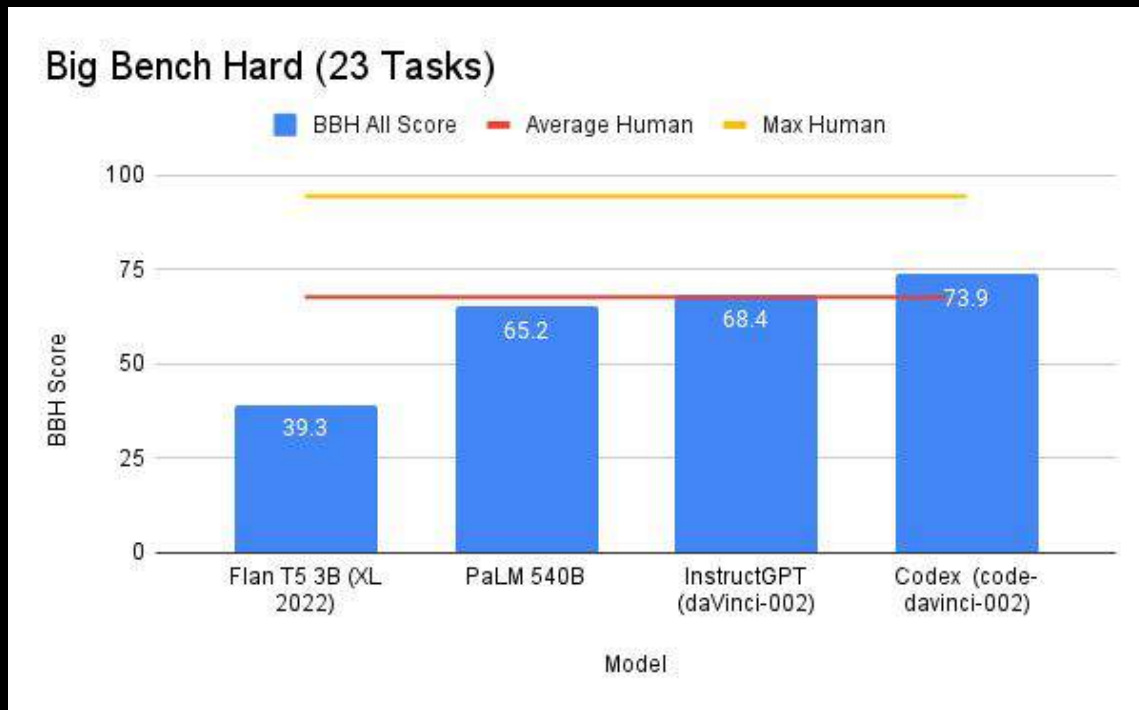
Rajiv Shah | @rajistics  
raj@hf.co



# Ability for LLMs to Perform Reasoning Tasks

Very good at solving reasoning tasks:

- Algorithmic/Arithmetic
- Natural Language Understanding
- Use of World Knowledge



<https://github.com/google/BIG-bench/>  
<https://arxiv.org/pdf/2210.09261.pdf>





# NLP Existing Use Cases

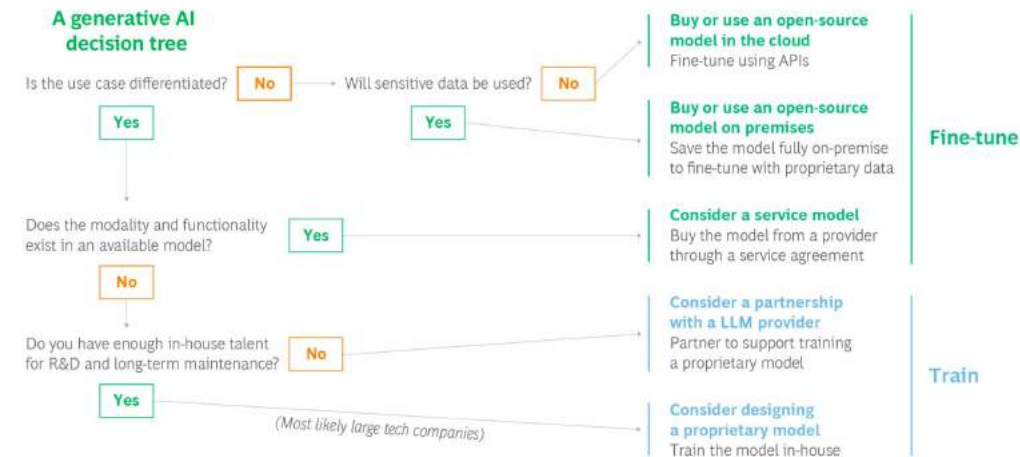
Differentiated?

Data Sensitivity?

Common Use

Case?

## Exhibit 1 - Choosing the Right Generative AI Model



Source: BCG Henderson Institute.

Note: API = application programming interface; LLM = large language model.



## Longer impact is going to be broader

- These LLMs can be the gateway to lots of other tools / services
- It's a way to 10X the value of your existing services by widening and easing access to them

