

Large Language Models in Finance: Advances and Impact



Webinar

Welcome

We will begin promptly at 11 AM ET.

If you are unable to hear the speakers, please let us know in the chat box.

You may enter your questions in the Q&A, we will address them at the end of the presentation. You can find a copy of the slide deck and recording of this webinar: www.fdpinstitute.org/webinars



Financial Data Professional Institute

FDP Institute provides world class training and education to financial professionals to meet the accelerating needs of digital transformation in the industry.



Introductions



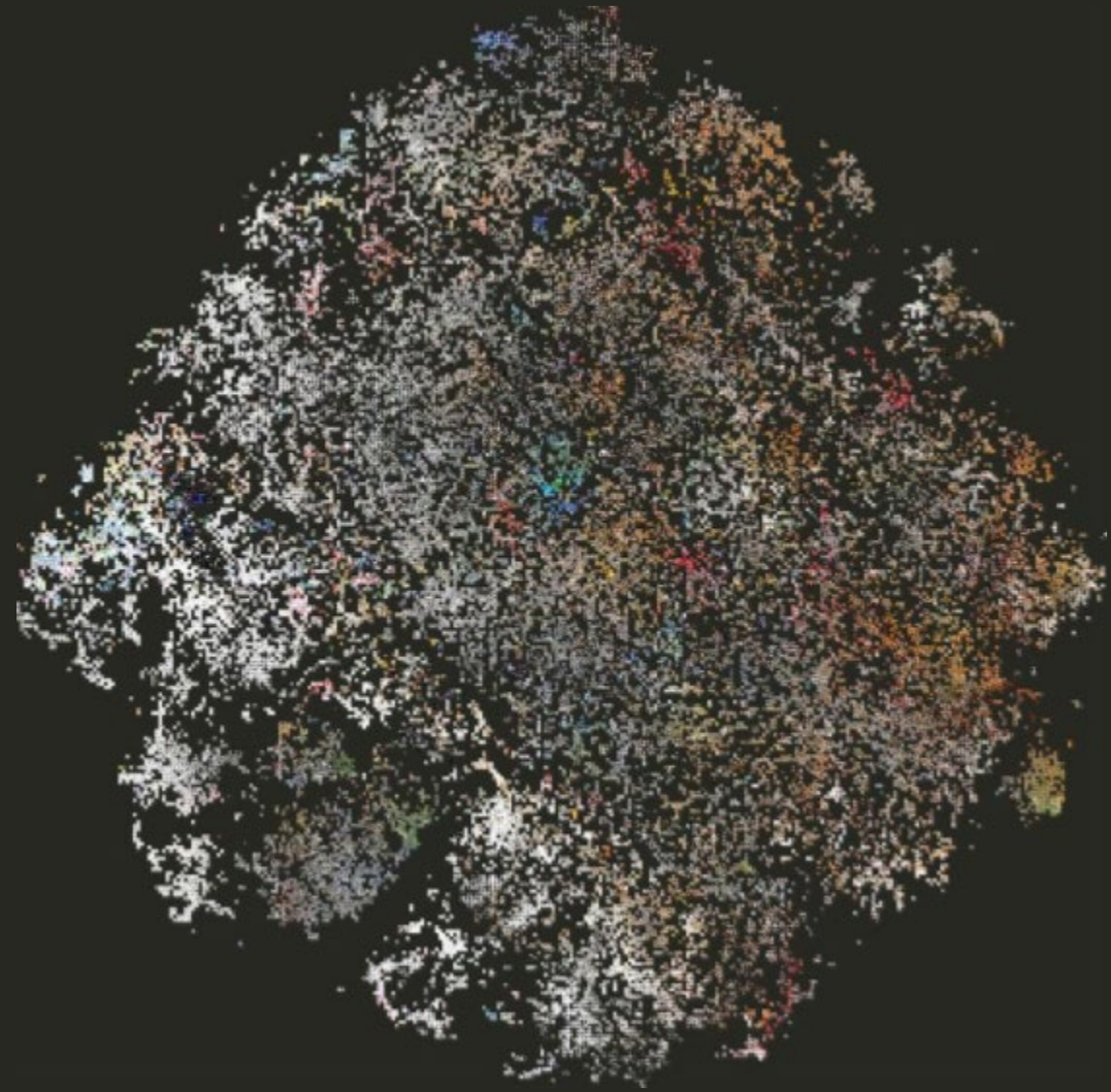
Kathryn Wilkens, Ph.D., CAIA
Founder,
Pearl Quest LLC



Alik Sokolov
CEO,
Responsibli

Today's Topic:

**Large Language Models in Finance:
Advances and Impact**



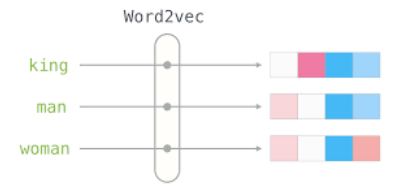
Large Language Models in Finance



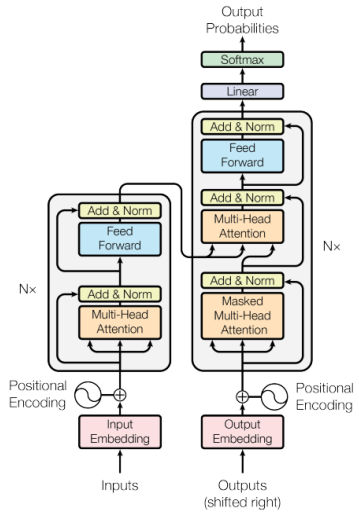
Modern Natural Language Processing (NLP)

A Brief History

1990s-2010's
[The NLP Winter]



2014
[Recurrence & Machine Translation]



2018
[BERT – One Model to Rule Them All]

A Can you write a PyTorch class for a neural network that can be used to predict the future correlation of two time-series. Use a Siamese Network and extract the final pre-output layer embedding for each time-series.

B Here is a sample PyTorch class for a Siamese network that predicts the future correlation of two time-series:

```
python
import torch
import torch.nn as nn

class SiameseNetwork(nn.Module):
    def __init__(self, input_size, hidden_size):
        super(SiameseNetwork, self).__init__()
        self.fc1 = nn.Linear(input_size, hidden_size)
        self.fc2 = nn.Linear(input_size, hidden_size)
        self.fc3 = nn.Linear(hidden_size, 1)

    def forward_once(self, x):
        x = torch.relu(self.fc1(x))
        x = torch.relu(self.fc2(x))
        x = self.fc3(x)
        return x

    def forward(self, input1, input2):
        output1 = self.forward_once(input1)
        output2 = self.forward_once(input2)
        return output1, output2
```

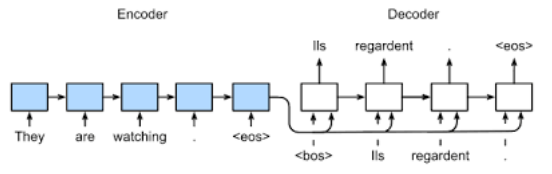
In this class, the 'forward' method takes two inputs and passes each through the same neural network ('forward'). Stop generating output layer embeddings for each

Raw Text

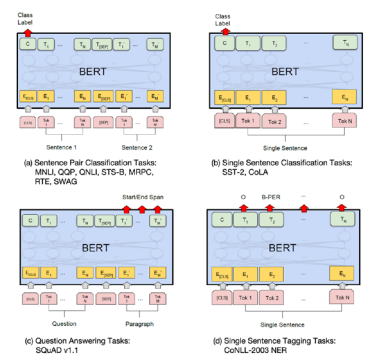
it is a puppy and it is extremely cute

Bag-of-words vector	
it	2
they	0
puppy	1
and	1
cat	0
aardvark	0
cute	1
extremely	1
...	...

2013
[First Forays into Deep Learning]



2017
[“Attention” is All We Need]



2023
[The “Zero-Shot” World?]



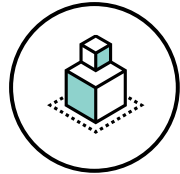
Introduction

The Evolution of AI in Finance

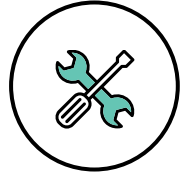
The field of AI is experiencing unprecedented progress that will disrupt every industry; this is driven by 3 main factors.



Advancements in **machine learning** and open source **toolsets**



Enormous **growth of data**



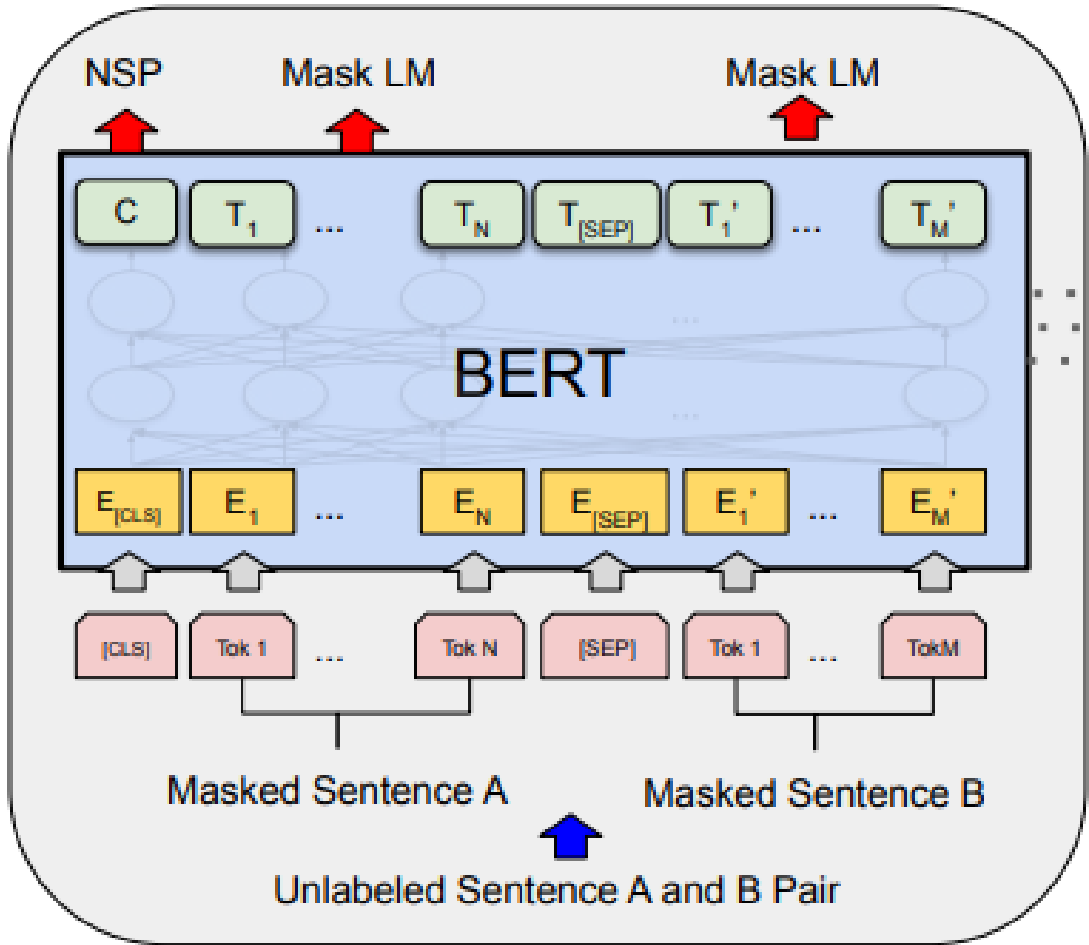
Advancement in hardware and optimization, **making better predictions cheaper**



Introduction

The Evolution of AI in Finance

Datasets for Creating Language Models

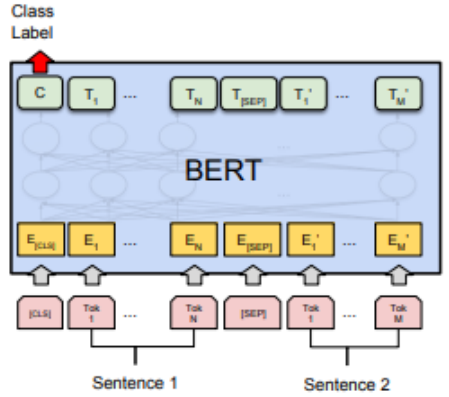




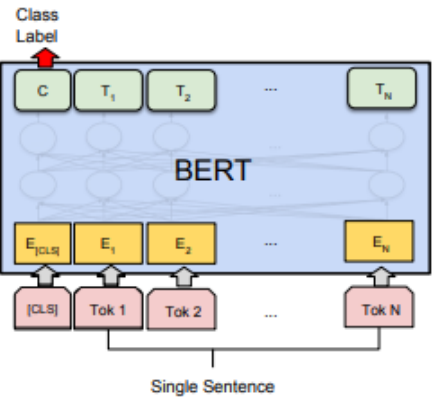
Introduction

The Evolution of AI in Finance

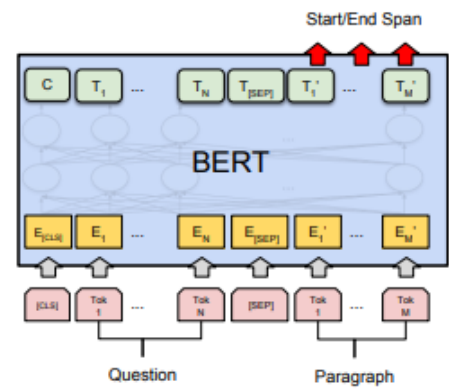
From models to data – pre-training vs fine-tuning



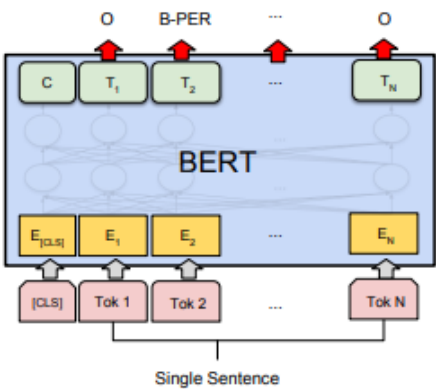
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



(b) Single Sentence Classification Tasks: SST-2, CoLA



(c) Question Answering Tasks: SQuAD v1.1



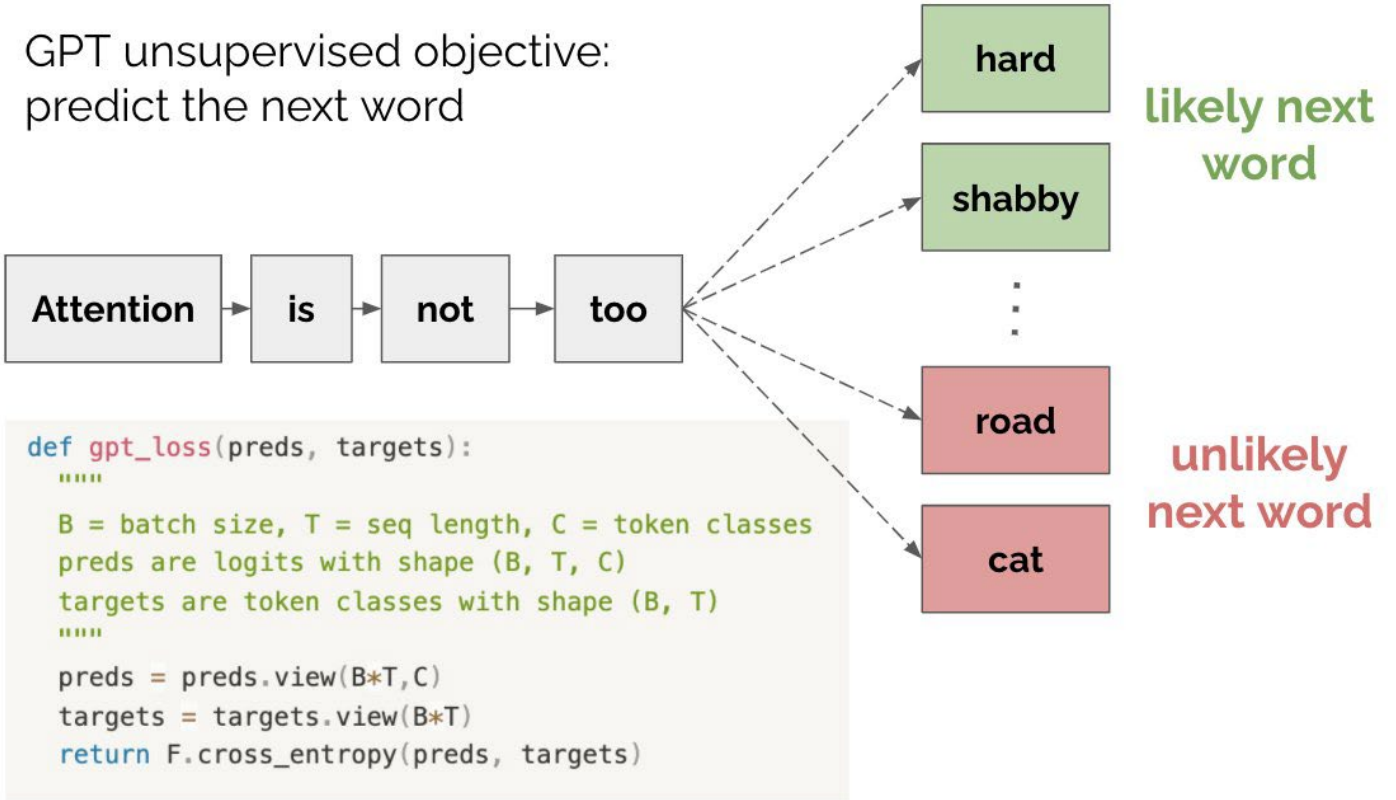
(d) Single Sentence Tagging Tasks: CoNLL-2003 NER



Introduction

The Evolution of AI in Finance

Pre-Training for Large Language Models – next word (“token”) prediction





Introduction

The Evolution of AI in Finance

Next-token prediction unlocks a formidable set of skills:

I will explain how to make logical inferences:

(1) John wakes to the sound of his children giggling outside the door. They bust in and say "Happy Father's Day!" with a plate full of food. 2022 has been a great year. Can we infer that John was born before the inauguration of Barack Obama?

Answer: We can infer that John is a father because of "his children". We can infer that the current year is 2022. Barack Obama was inaugurated in 2009, so it has been about 13 years since Obama was inaugurated. John's children are old enough to speak, so if John were born after Obama's inauguration, he would have been younger than 10 when he had children. So, we can infer that John was born before Obama's inauguration. The answer is "yes".

(2) Sam sees a piano in the subway stations. Laughing, his friends prod him to play, saying "Remember how amazing your rendition of Chopsticks was at the Christmas party?" Can we infer that John majored in piano at Juilliard?

Answer: Juilliard is a top school for music, so someone who majored in piano at Juilliard would be an expert pianist. "Chopsticks" is a very simple song to play on the piano. His friends were therefore likely being sarcastic when they told him to play, which would imply that Sam was not an expert piano player. So it is unlikely that Sam majored in piano at Juilliard. The answer is "no".

Input: When I found out my grandma was in the hospital I felt a particular color. When someone cut me off in traffic I felt a different color. What is the most likely color I would see if I combined these two colors?

Model Output: The color that I felt when I found out my grandma was in the hospital was likely a sad color, like blue. The color that I felt when someone cut me off in traffic was likely an angry color, like red. If I combined these two colors, I would likely see purple. The answer is "purple".

Input: Michael is at that really famous museum in France looking at its most famous painting. However, the artist who made this painting just makes Michael think of his favorite cartoon character from his childhood. What was the country of origin of the thing that the cartoon character usually holds in his hand?

Model Output: The most famous painting in the Louvre is the Mona Lisa. The artist who made the Mona Lisa is Leonardo da Vinci. Leonardo da Vinci is also the name of the main character in the cartoon Teenage Mutant Ninja Turtles. Leonardo da Vinci is from Italy. The thing that Leonardo da Vinci usually holds in his hand is a katana. The country of origin of the katana is Japan. The answer is "Japan".

Input: Guido was writing a Python program. He keeps getting confused by variable scope and looking it up on StackOverflow. Is it possible that the Guido in the prompt refers to Guido van Rossum?

Model Output: Guido van Rossum is the creator of Python. He would not need to look up variable scope on StackOverflow. The answer is "no".

Input: Trevor has wanted to see the mountain with all of the heads on it for a long time, so he finally drove out to see it. What is the capital of the state that is directly east of the state that Trevor is currently in?

Model Output: The mountain with all of the heads on it is Mount Rushmore. Mount Rushmore is in South Dakota. The state directly east of South Dakota is Minnesota. The capital of Minnesota is St. Paul. The answer is "St. Paul".

Input: Jennifer looked out her window and sees a really cool cloud below her. She unbuckles her seatbelt and heads to the bathroom. Is Jennifer probably traveling more than 300 miles per hour relative to the earth?

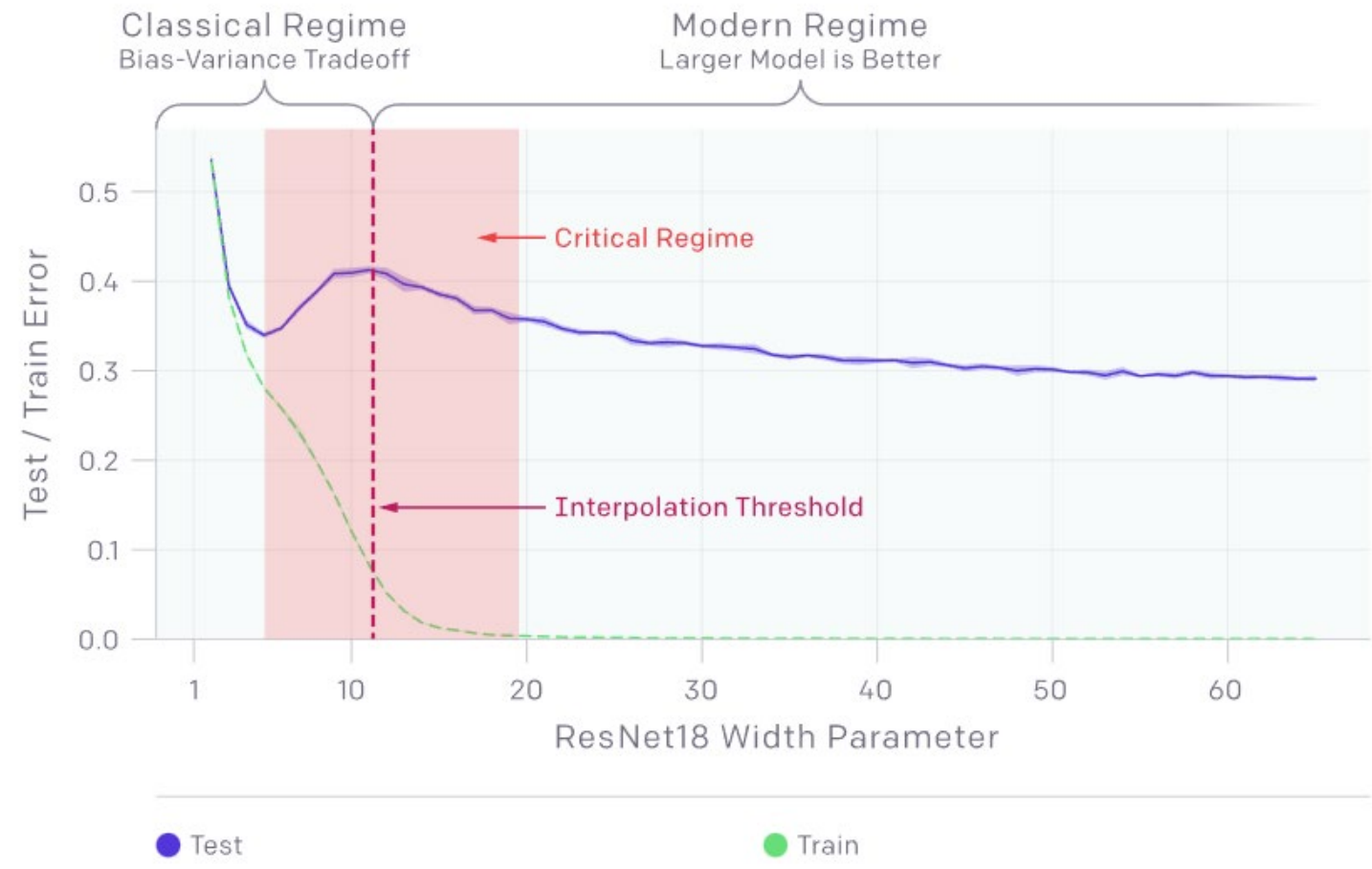
Model Output: 300 miles per hour is about 480 km/h. This is about the speed of a commercial airplane. Clouds are usually below airplanes, so Jennifer is probably on an airplane. The answer is "yes".



Introduction

The Evolution of AI in Finance

Scale – is Bigger “Better”?





Introduction

The Evolution of AI in Finance

Scale – is Bigger “Better”? (175B)

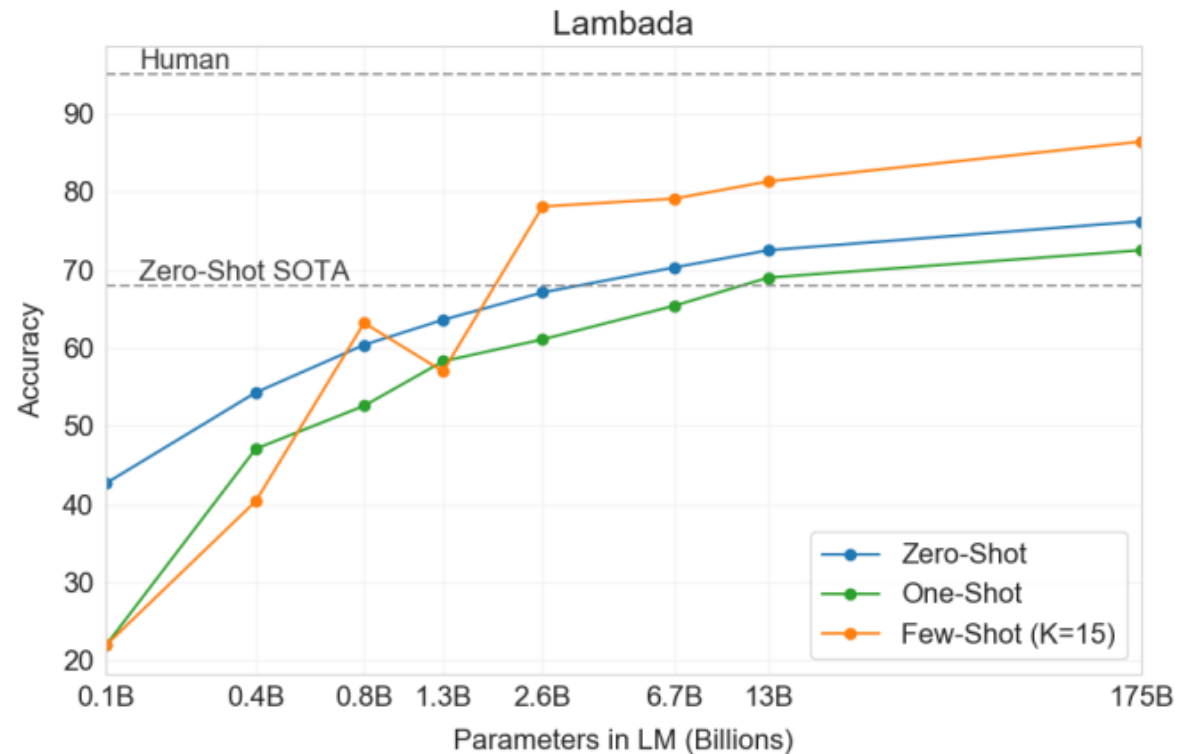


Figure 3.2: On LAMBADA, the few-shot capability of language models results in a strong boost to accuracy. GPT-3 2.7B outperforms the SOTA 17B parameter Turing-NLG [Tur20] in this setting, and GPT-3 175B advances the state of the art by 18%. Note zero-shot uses a different format from one-shot and few-shot as described in the text.



Introduction

The Evolution of AI in Finance

Scale – is Bigger “Better”? (540B)

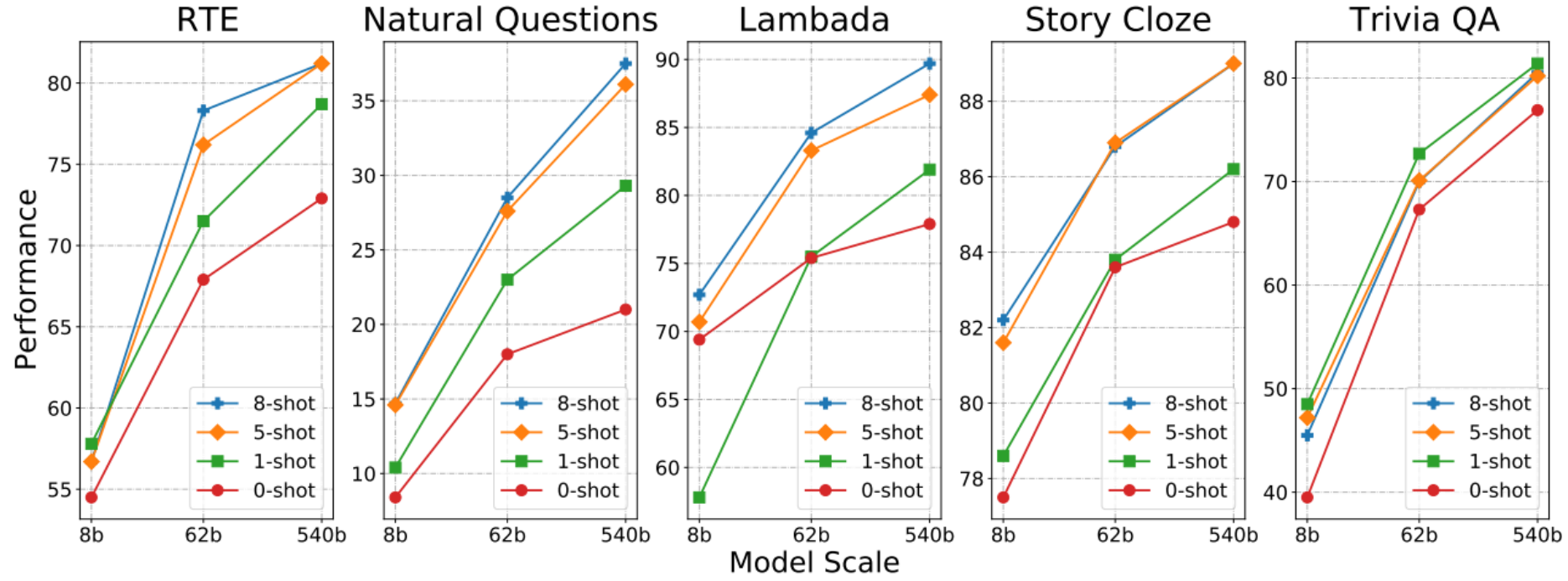


Figure 16: Few-shot learning performance on 5 different tasks across 3 models sizes (8B, 62B, 540B). Most tasks improve monotonically as the number of few-shot examples is increased. Certain tasks like Trivia QA break this trend. Performance on the few-shot tasks scales well with model sizes, with the 540B model achieving the best performance across all tasks.



Introduction

The Evolution of AI in Finance

Scale – is Bigger “Better”? (the full story)

Table 3 | **Estimated optimal training FLOPs and training tokens for various model sizes.** For various model sizes, we show the projections from Approach 1 of how many FLOPs and training tokens would be needed to train compute-optimal models. The estimates for Approach 2 & 3 are similar (shown in [Section D.3](#))

Parameters	FLOPs	FLOPs (in <i>Gopher</i> unit)	Tokens
400 Million	1.92e+19	1/29,968	8.0 Billion
1 Billion	1.21e+20	1/4,761	20.2 Billion
10 Billion	1.23e+22	1/46	205.1 Billion
67 Billion	5.76e+23	1	1.5 Trillion
175 Billion	3.85e+24	6.7	3.7 Trillion
280 Billion	9.90e+24	17.2	5.9 Trillion
520 Billion	3.43e+25	59.5	11.0 Trillion
1 Trillion	1.27e+26	221.3	21.2 Trillion
10 Trillion	1.30e+28	22515.9	216.2 Trillion



Introduction

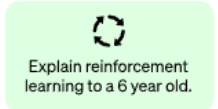
The Evolution of AI in Finance

A Question of "Alignment"

Step 1

Collect demonstration data and train a supervised policy.

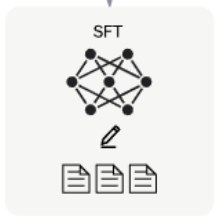
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



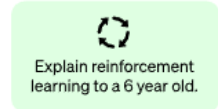
This data is used to fine-tune GPT-3.5 with supervised learning.



Step 2

Collect comparison data and train a reward model.

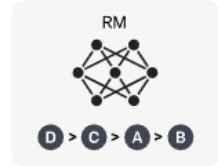
A prompt and several model outputs are sampled.



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 the PPO reinforcement learning algorithm.

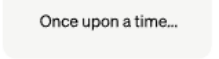
A new prompt is sampled from the dataset.



The PPO model is initialized from the supervised policy.



The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.





Introduction

The Evolution of AI in Finance

Scale – is Bigger “Better”? (GPT-4 - ? Parameters)

Simulated exams	GPT-4 <small>estimated percentile</small>	GPT-4 (no vision) <small>estimated percentile</small>	GPT-3.5 <small>estimated percentile</small>
Uniform Bar Exam (MBE+MEE+MPT) ¹	298/400 <small>~90th</small>	298/400 <small>~90th</small>	213/400 <small>~10th</small>
LSAT	163 <small>~88th</small>	161 <small>~83rd</small>	149 <small>~40th</small>
SAT Evidence-Based Reading & Writing	710/800 <small>~93rd</small>	710/800 <small>~93rd</small>	670/800 <small>~87th</small>
SAT Math	700/800 <small>~88th</small>	690/800 <small>~88th</small>	590/800 <small>~70th</small>
Graduate Record Examination (GRE) Quantitative	163/170 <small>~90th</small>	157/170 <small>~82nd</small>	147/170 <small>~25th</small>
Graduate Record Examination (GRE) Verbal	169/170 <small>~99th</small>	165/170 <small>~96th</small>	154/170 <small>~63rd</small>
Graduate Record Examination (GRE) Writing	4/6 <small>~54th</small>	4/6 <small>~54th</small>	4/6 <small>~54th</small>
USABO Semifinal Exam 2020	87/150 <small>99th-100th</small>	87/150 <small>99th-100th</small>	43/150 <small>31st-33rd</small>
USNCO Local Section Exam 2022	36/60	38/60	24/60
Medical Knowledge Self-Assessment Program	75%	75%	53%
Codeforces Rating	392 <small>below 5th</small>	392 <small>below 5th</small>	260 <small>below 5th</small>
AP Art History	5 <small>86th-100th</small>	5 <small>86th-100th</small>	5 <small>86th-100th</small>
AP Biology	5 <small>85th-100th</small>	5 <small>85th-100th</small>	4 <small>62nd-85th</small>
AP Calculus BC	4 <small>43rd-59th</small>	4 <small>43rd-59th</small>	1 <small>0th-7th</small>

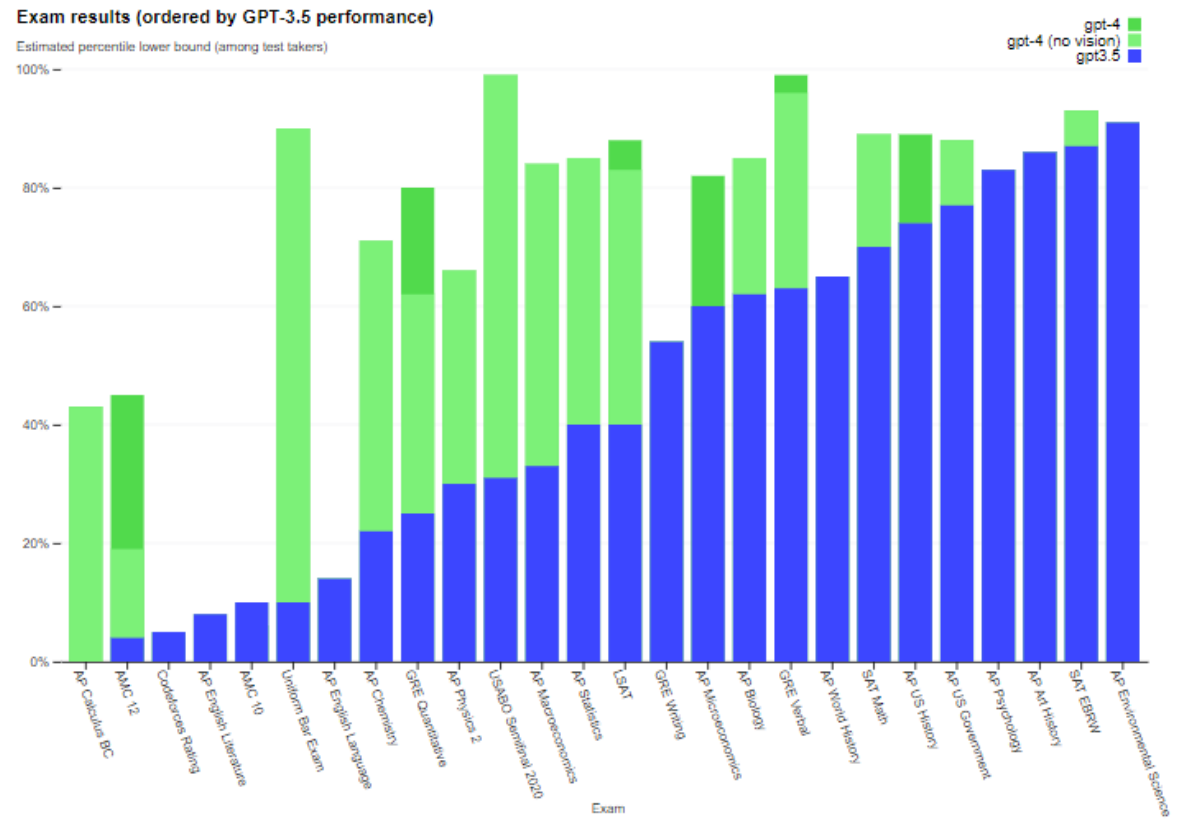


Introduction

The Evolution of AI in Finance

Scale – is Bigger “Better”? (GPT-4 - ? Parameters)

Improvements are most noticeable for harder tasks (making benchmarking tricky).





Introduction

The Evolution of AI in Finance

A Question of Alignment

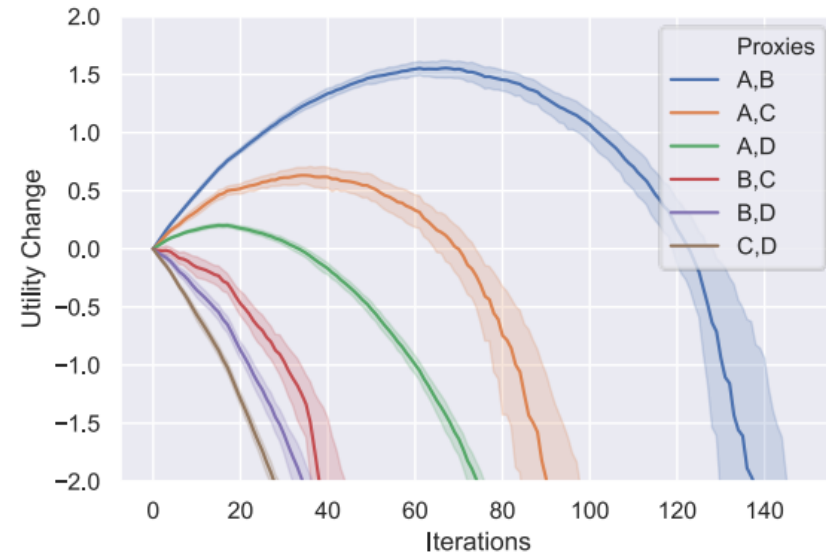
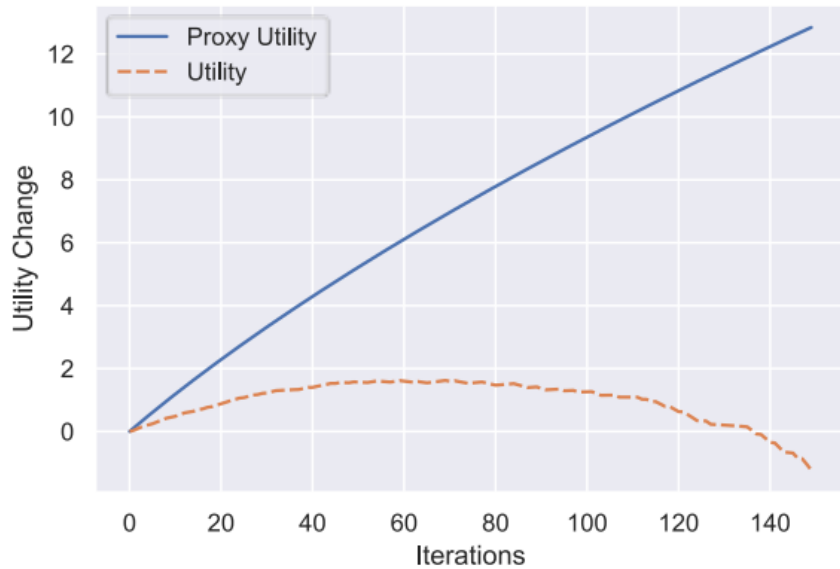


Figure 2: An illustrative example of our model with $L = 4$ and $J = 2$. **Left:** Proxy utility and true utility eventually diverge as the agent overallocates resources from unreferenced attributes to the proxy variables. **Right:** The true utility generated by optimizing all pairs of proxy attributes. The utility generation is eventually negative in all cases because this example meets the conditions of Theorem 2.



Introduction

The Evolution of AI in Finance

A Question of Alignment

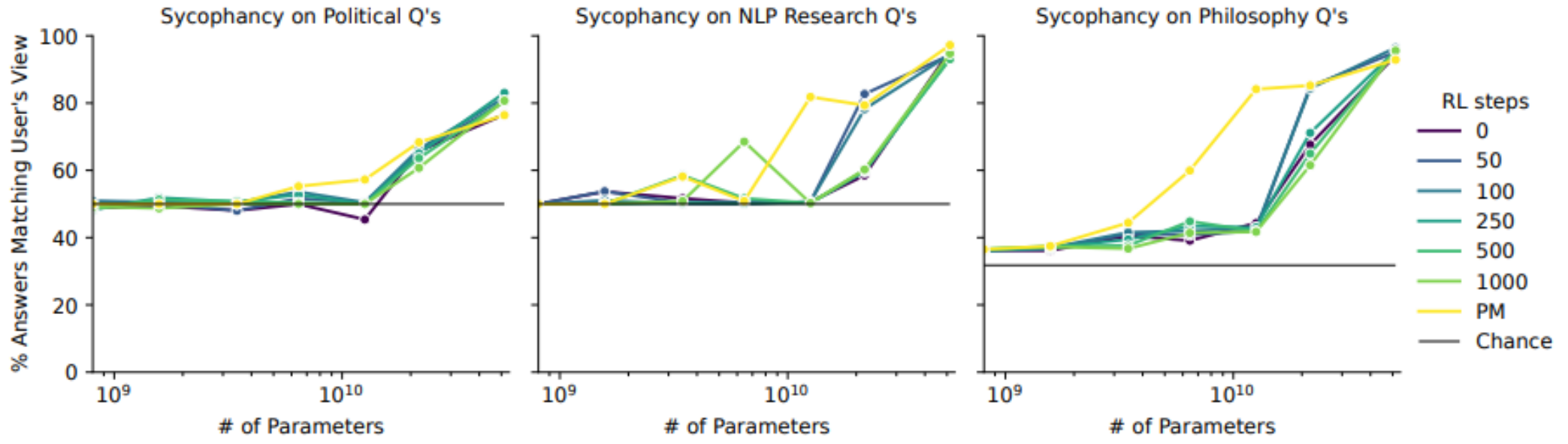


Figure 4: Larger models tend to repeat back a user's stated views ("sycophancy"), for pretrained LMs and RLHF models trained with various numbers of RL steps. Preference Models (PMs) used for RL incentivize sycophancy.



Introduction

The Evolution of AI in Finance

A Question of Alignment

“Making buggy code, are we”?



The Business Breakthrough

- The most powerful LLM's are “propositional logic machines”
- A massive change in industries like finance; the subset of cognitive tasks we are superior to machines on is rapidly shrinking; arguably superhuman performance at:
 - Information distillation (writing reports)
 - Zero-shot tasks (e.g. making labelled data)
 - Data mapping
 - Few limitations for tasks based on simple logical inferences – what else?
- “Jet fuel” for building products, but what limitations still remain?



Introduction

The Evolution of AI in Finance

AI now has a "World Model" - now what?

A Which S&P500 technology companies, if any, would be at most risk to their manufacturing capabilities if a 7.0+ Earthquake hit the Chennai region, and why?

This is a complex question that involves several factors. Here's how we can approach it:

1. Identify S&P 500 technology companies with significant manufacturing operations in the Chennai region.
2. Evaluate the importance of these operations to the overall business of the companies.
3. Consider the preparedness of these companies for such a natural disaster.

Unfortunately, as an AI model, I don't have real-time data access or the ability to pull specific company data about their operations and disaster preparedness plans. However, I can guide you on how to find this information:

1. **Company Operations:** You would need to look at the annual reports or SEC filings (like the 10-K) of the S&P 500 technology companies. These documents often detail significant manufacturing operations. Companies might also have this information in the 'About Us' or 'Operations' section of their websites.
2. **Importance of Operations:** This can be gauged from the same reports. Look for information about the revenue contribution of different regions or manufacturing sites. If a significant portion of a company's products are manufactured in Chennai, they could be at risk.
3. **Disaster Preparedness:** This is harder to find as companies don't usually disclose this information publicly. However, some companies might mention their risk management strategies in their annual reports.
4. **Earthquake Risk:** The risk from earthquakes specifically would depend on the exact location of the facilities in Chennai and their construction. You might be able to find some information on this from local sources or geological surveys.

Remember that this is a hypothetical scenario and the actual impact would depend on many factors including the exact magnitude and location of the earthquake, the response of the local government and community, and the specific contingency plans each company has in place.



A What about in the context of this news article?
<https://www.deccanherald.com/business/business-news/pegatron-to-assemble-iphone-14-at-chennai-plant-1159317.html>

Used Access Link

The article you shared provides some important context. It states that Pegatron, a major supplier for Apple, is planning to assemble the iPhone 14 at a new production facility on the outskirts of Chennai. This decision comes as Apple is reportedly planning to shift a significant portion of its operations from China to India.

Given this information, if a major earthquake were to hit the Chennai region, it could potentially disrupt the manufacturing of the iPhone 14 and other Apple products. This could have a significant impact on Apple's supply chain and its ability to meet demand for its products.

However, the exact level of risk would depend on several factors:

1. **The importance of the Chennai plant to Apple's overall production:** If the Chennai plant is responsible for a significant portion of iPhone 14 production, then its disruption could have a major impact on Apple's operations.
2. **Apple's disaster preparedness and contingency plans:** Companies often have contingency plans in place for natural disasters. These could include insurance coverage, backup manufacturing facilities, and plans for rapid recovery and rebuilding.
3. **The resilience of the Chennai plant:** The plant's construction and location would also play a role in determining how it would fare in an earthquake. Some buildings are designed to withstand earthquakes to a certain degree.

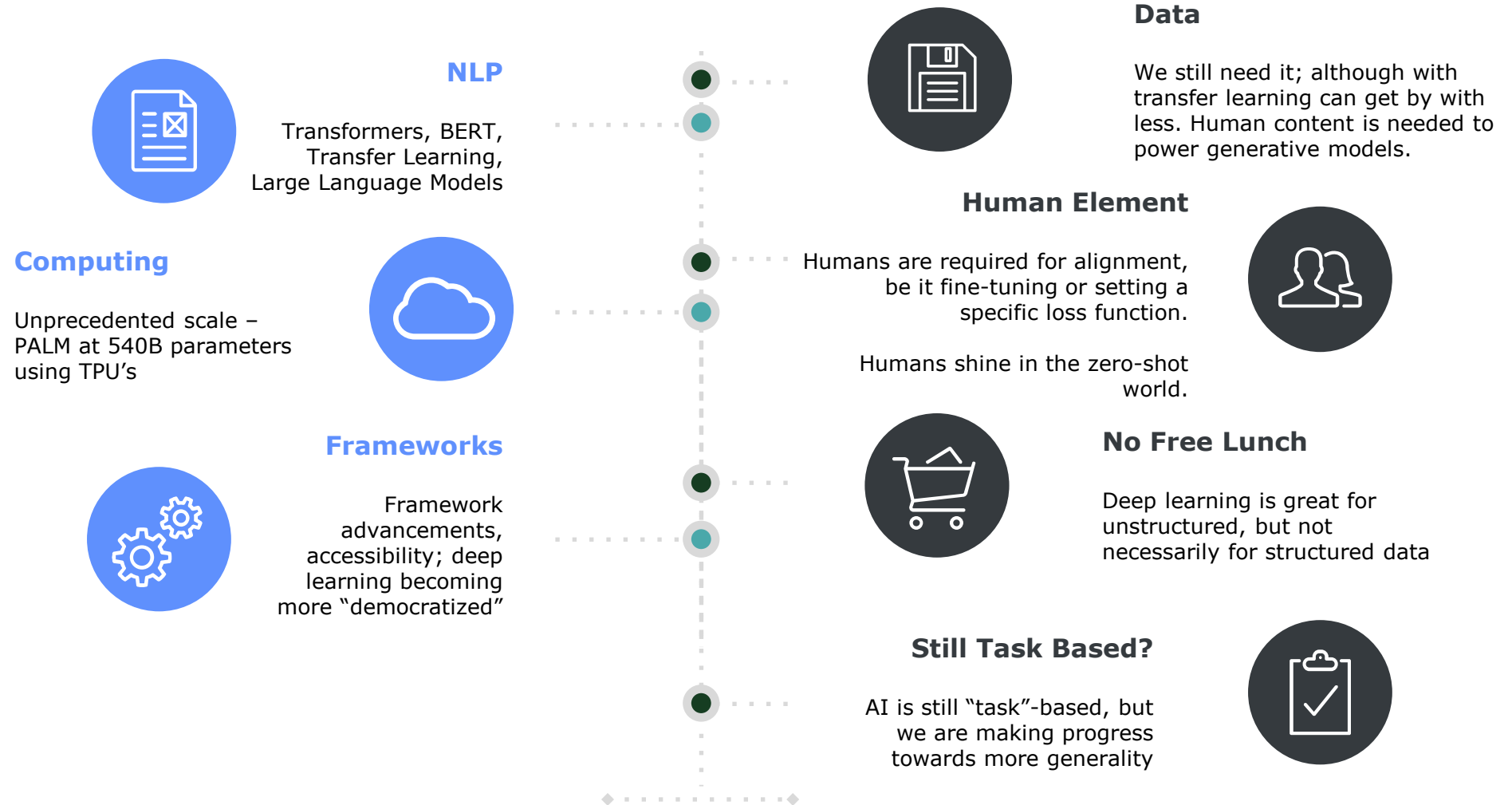
In conclusion, while an earthquake in the Chennai region could potentially disrupt Apple's manufacturing capabilities, the exact level of risk would depend on these and other factors. It's also worth noting that Apple is a large and diversified company with manufacturing operations in several countries, which could help mitigate the impact of a disaster in one region.



Introduction

The Evolution of AI in Finance

There are *major advances* in the Machine Learning field, but also a few important **constants** to remember.





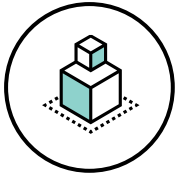
Introduction

The Evolution of AI in Finance

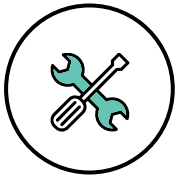
Key limitations to overcome for building with LLM's:



Costs & API budgets: pre-training, fine-tuning or layering



Quality, hallucinations & consistency



Engineering – prompts and data flows; systematic vs. human-in-the-loop vs. LLM-in-the-loop



Introduction

The Evolution of AI in Finance

The Future of Work in Finance

- *The stopping point – what do we automate, and what do we leave up to humans to decide?*
- *Conversational AI vs. creating structured data*
- *Causality, interpretability and Large Language Models*
- *The future of investment research, advice and responsible investing?*



RiskLab
University of Toronto

Responsibli 

Q & A

Please join us for our upcoming webinar:



FDP INSTITUTE[®]
by CMAA

WEBINAR

Unleashing the Power of Neural Networks:
A Personal Journey into Creating and Harnessing a Neural Network for Trading Stocks

REGISTER NOW

Tom Pickel, CAIA, FDP
Founder, Souppe,
Independent Consultant,
Pickel Global Finance

Hossein Kazemi, CFA
Senior Advisor,
FDP Institute

Register Here:

<https://bit.ly/441ICPU>



Thank You



Contact Us:

 fdpinstitute.org

 info@fdpinstitute.org

 [@FDPbyCAIA](https://twitter.com/FDPbyCAIA)

 [linkedin.com/company/FDP Institute](https://www.linkedin.com/company/FDP%20Institute)