

Artificial Intelligence for Business Research @Antai

# Deep-Learning-based NLP: Pretraining and Large Language Models

Renyu (Philip) Zhang

1

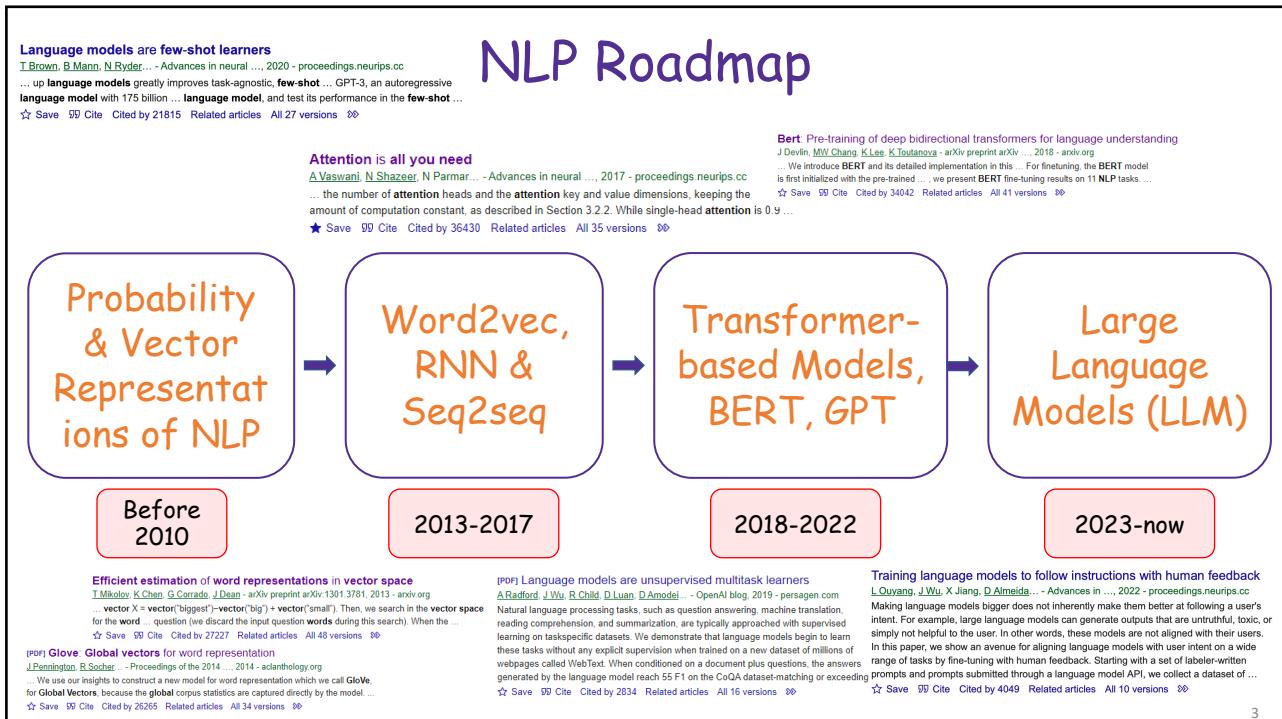
## Agenda

- Pre-training: BERT & GPT
- Large Language Models

2

2

1



3

**Jason Wei's Typical Day**

My typical day as a Member of Technical Staff at OpenAI:  
[9:00am] Wake up  
[9:30am] Commute to Mission SF via Waymo. Grab avocado toast from Tartine  
[9:45 am] Recite OpenAI charter. Pray to optimization Gods. Learn the Bitter Lesson  
[10:00am] Meetings (Google Meet). Discuss how to train larger models on more data  
[11:00am] Write code to train larger models on more data. pair= @hwchung27

What is The Bitter Lesson?

4

## The Bitter Lesson



- Reference: <http://www.incompleteideas.net/IncIdeas/BitterLesson.html>
- The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin.
- Leveraging domain knowledge (short-term & specific) vs. Leveraging computation (long-term & general).
- Bitter lesson: Leveraging domain knowledge is self-satisfying and intellectually inspiring, but plateaus in the long-run or even inhibits further progress.

5

5

## Jason Wei's Typical Day (Cont'd)

[12:00pm] Lunch at the canteen (vegan, gluten-free)  
 [1:00pm] Actually train large models models on more data  
 [2:00pm] Debug infra issues (why the fck did I pull from master?)  
 [3:00pm] Babysit model training. Play with Sora  
 [4:00pm] Prompt engineer aforementioned large models trained on more data  
 [4:30pm] Short break, sit on avocado chair. Wonder how good Gemini Ultra actually is  
 [5:00pm] Brainstorm potential algorithmic improvements for models  
 [5:05pm] Conclude that algorithmic changes are too risky. Safer to just scale compute and data  
 [6:00pm] Dinner. Clam chowder with Roon  
 [7:00pm] Commute back home  
 [8:00pm] Have a wine and get back to coding. Ballmer's peak is coming  
 [9:00pm] Analyze experimental runs. I have a love/hate relationship with wandb  
 [10:00pm] Launch experiments to run overnight and get results by tomorrow morning  
 [1:00am] Experiments actually get launched  
 [1:15am] Bedtime. Satya and Jensen watch from above. Compression is all you need. Good night

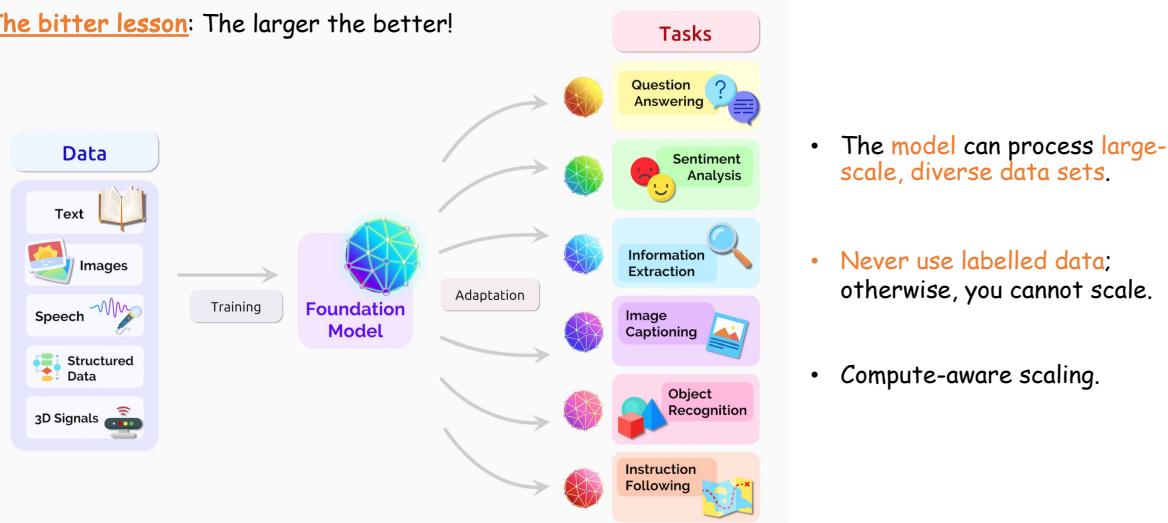
6

6

## Pretraining: Scaling Unsupervised Learning on the Internet

References: Stanford CS224N, Lecture 9: <https://web.stanford.edu/class/cs224n/slides/cs224n-2024-lecture09-pretraining.pdf>  
 Princeton COS 597G, Lecture 2: <https://www.cs.princeton.edu/courses/archive/fall22/cos597G/lectures/lec02.pdf>

The bitter lesson: The larger the better!



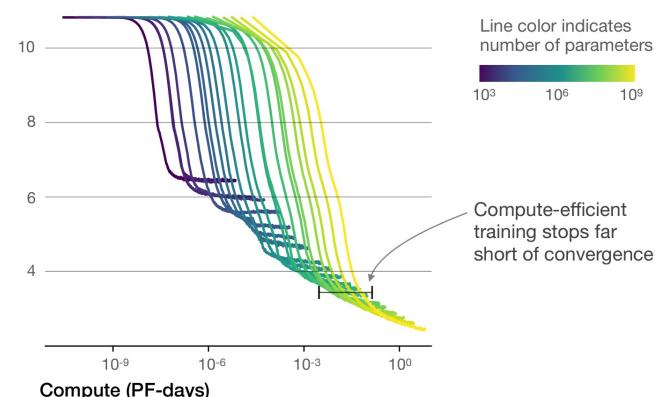
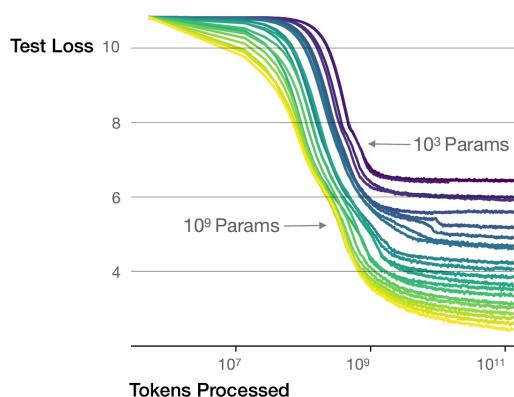
7

7

## Scaling Laws

Larger models require **fewer samples** to reach the same performance

The optimal model size grows smoothly with the loss target and compute budget



### Scaling laws for neural language models

J.Kaplan, S.McCandlish, T.Henighan, I.B.Brown... - arXiv preprint arXiv ..., 2020 - arxiv.org

... scaling laws for language model performance on the cross-entropy loss. The loss scales as a power-law with model ... Simple equations govern the dependence of overfitting on model/...

☆ Save 99 Cite Cited by 965 Related articles All 3 versions ☺

### Training compute-optimal large language models

J.Hoffmann, S.Borgeaud, A.Mensch... - arXiv preprint arXiv ..., 2022 - arxiv.org

We investigate the optimal model size and number of tokens for training a transformer language model under a given compute budget. We find that current large language models are ...

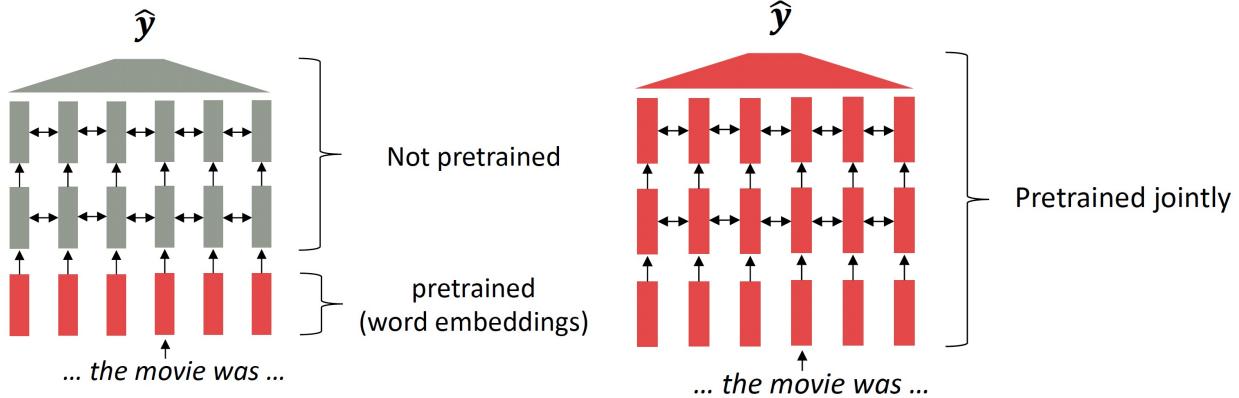
☆ 保存 99 引用 被引用次数: 942 相关文章 所有 6 个版本 ☺

8

8

## From Pretrained Word Embeddings to Pretrained Models

References: Stanford CS224N, Lecture 9: <https://web.stanford.edu/class/cs224n/slides/cs224n-2024-lecture09-pretraining.pdf>  
 Princeton COS 597G, Lecture 2: <https://www.cs.princeton.edu/courses/archive/fall22/cos597G/lectures/lec02.pdf>



[Recall, *movie* gets the same word embedding,  
no matter what sentence it shows up in]

[This model has learned how to represent  
entire sentences through pretraining]

9

9

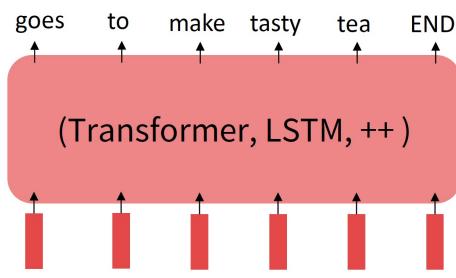
## From Pretrained Word Embeddings to Pretrained Models

References: Stanford CS224N, Lecture 9: <https://web.stanford.edu/class/cs224n/slides/cs224n-2024-lecture09-pretraining.pdf>  
 Princeton COS 597G, Lecture 2: <https://www.cs.princeton.edu/courses/archive/fall22/cos597G/lectures/lec02.pdf>

- Pretraining can improve downstream NLP applications by serving as **parameter initialization**.

### Step 1: Pretrain (on language modeling)

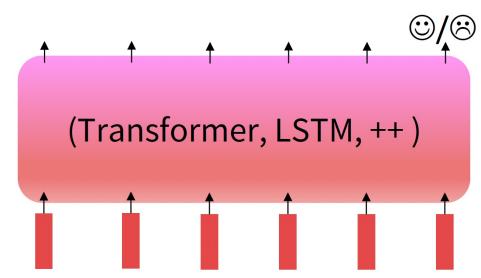
Lots of text; learn general things!



$\hat{\theta}$  by approximating  $\min_{\theta} \mathcal{L}_{\text{pretrain}}(\theta)$

### Step 2: Finetune (on your task)

Not many labels; adapt to the task!



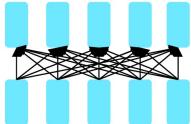
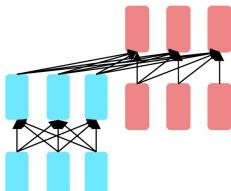
approximates  $\min_{\theta} \mathcal{L}_{\text{finetune}}(\theta)$ , starting at  $\hat{\theta}$

10

10

## Three Pretraining Architectures

References: Stanford CS224N, Lecture 9: <https://web.stanford.edu/class/cs224n/slides/cs224n-2024-lecture09-pretraining.pdf>  
 Princeton COS 597G, Lecture 2: <https://www.cs.princeton.edu/courses/archive/fall22/cos597G/lectures/lec02.pdf>

 <b>Encoders</b>	<ul style="list-style-type: none"> <li>• Can condition on future.</li> <li>• Example: BERT.</li> </ul>
 <b>Encoder-Decoders</b>	<ul style="list-style-type: none"> <li>• Combining encoder and decoder.</li> <li>• Example: T5</li> </ul>
 <b>Decoders</b>	<ul style="list-style-type: none"> <li>• Cannot condition on future.</li> <li>• Example: GPT</li> <li>• All (very) large language models are decoders.</li> </ul>

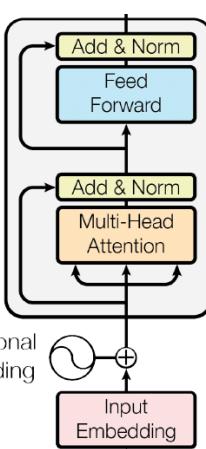
11

11

## BERT: Bidirectional Encoder Representations from Transformers

References: Stanford CS224N, Lecture 9: <https://web.stanford.edu/class/cs224n/slides/cs224n-2024-lecture09-pretraining.pdf>  
 Princeton COS 597G, Lecture 2: <https://www.cs.princeton.edu/courses/archive/fall22/cos597G/lectures/lec02.pdf>





N<sub>x</sub>  
N=12  
or 24

Positional Encoding

Input Embedding

Inputs

- Key idea: Learn representations based on **bidirectional context**.
  - We went to the **bank** of the river. vs. I need to go to the **bank** to make a deposit.
- Pretraining objectives: **masked language modeling** + **next sentence prediction**
- 15% of tokens are randomly masked. [Predict these!]
- The masked tokens in the inputs:
  - 80% replaced with [MASK];
  - 10% replaced with a random token;
  - 10% no change.
- Why not all masked tokens replaced with [MASK]?
- [MASK] tokens are never seen in fine-tuning.

Bert: Pre-training of deep **bidirectional** transformers for language understanding  
 J Devlin, MW Chang, K Lee, K Toutanova - arXiv preprint arXiv ..., 2018 - arxiv.org  
 ... BERT, which stands for **Bidirectional Encoder Representations** from Transformers. Unlike ...  
 2018), BERT is designed to pretrain deep **bidirectional representations** from unlabeled text by ...

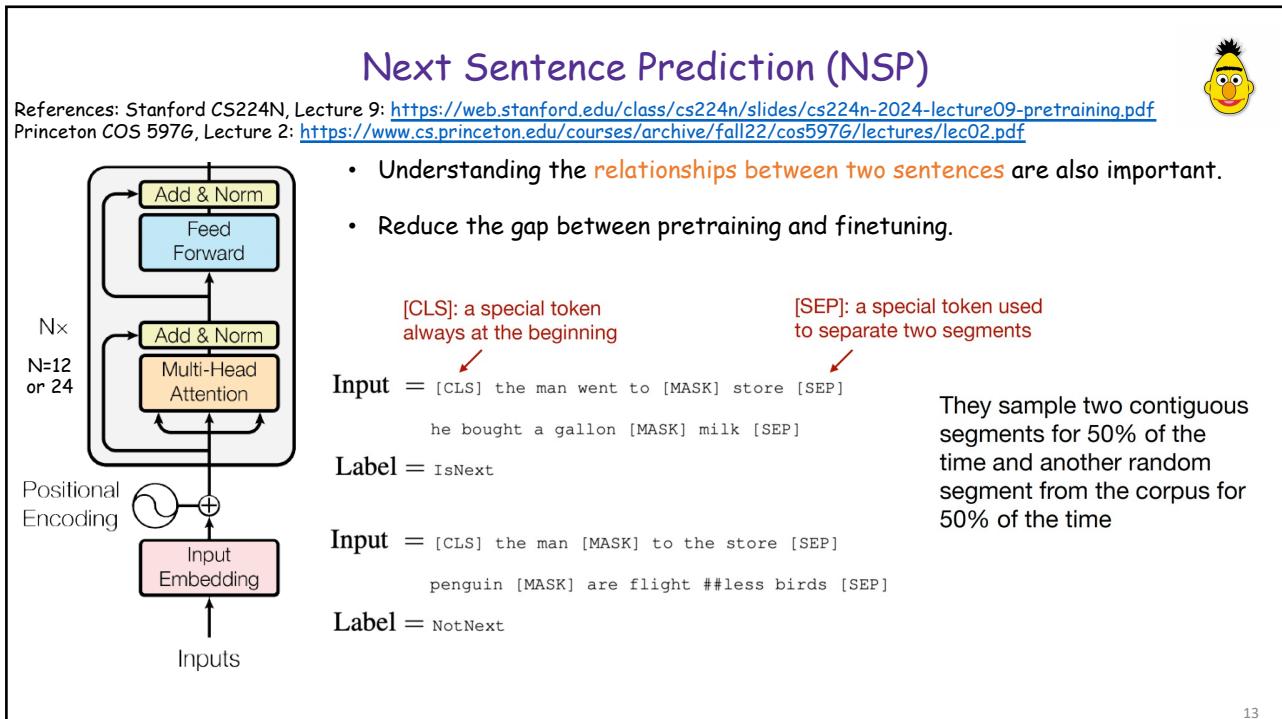
[Replaced] [Not replaced] [Masked]

Transformer Encoder

I pizza to the [M]

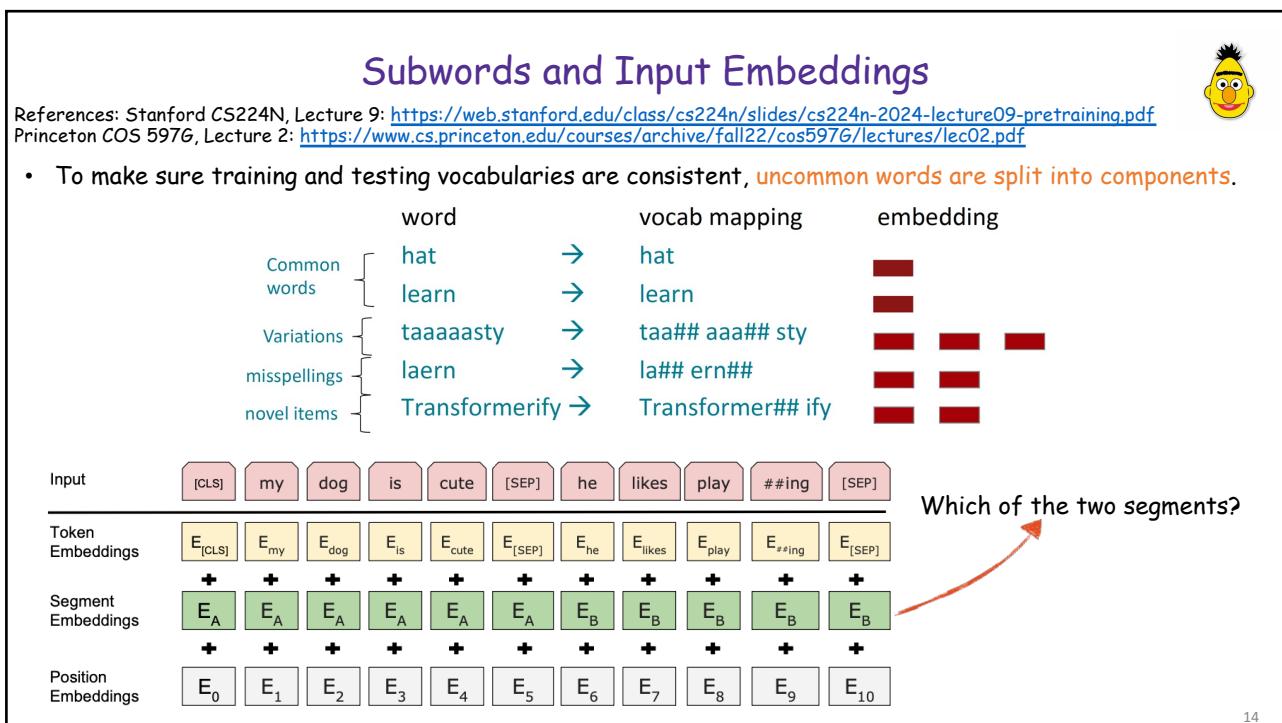
12

12



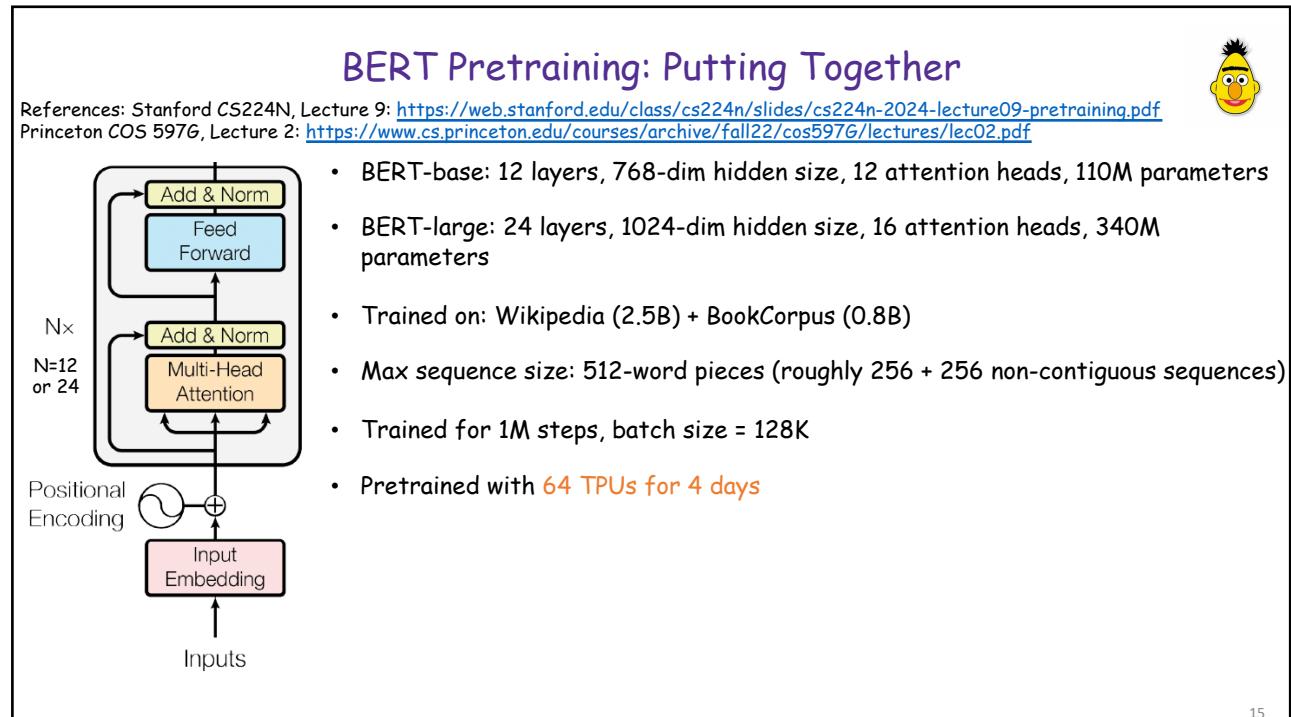
13

13



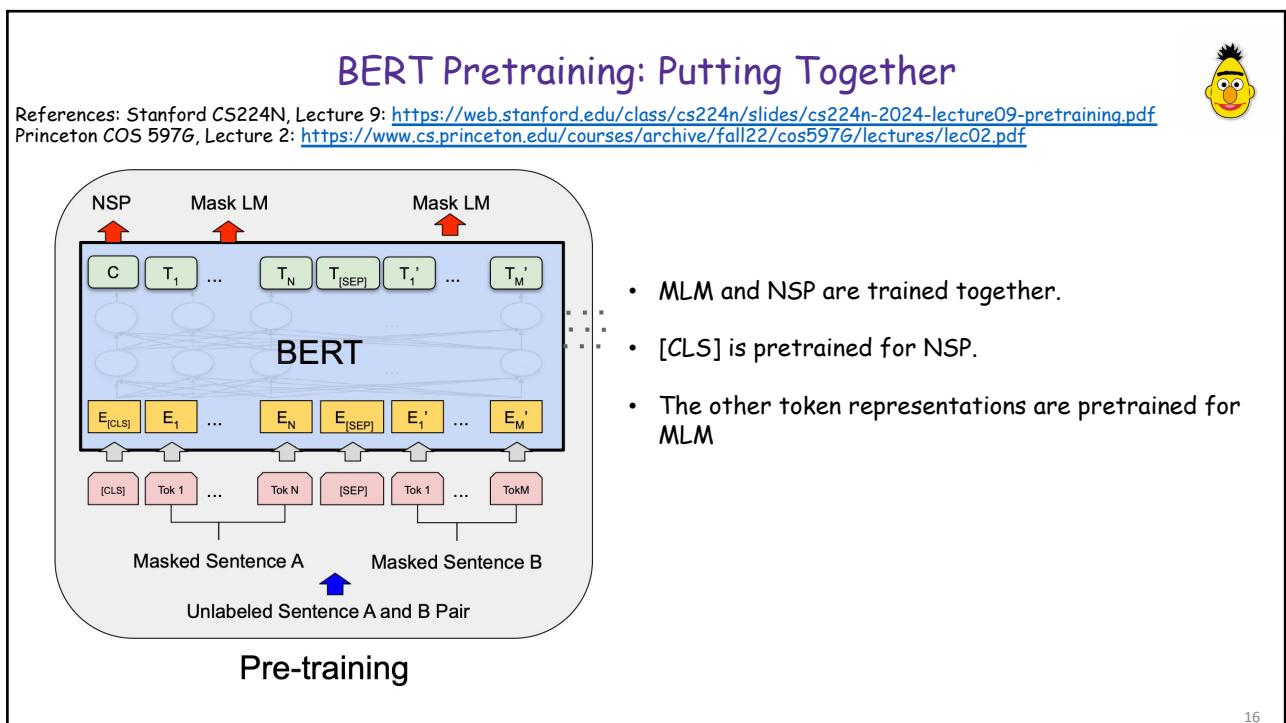
14

14



15

15



16

16

## Pretrain Once, Finetune Many Times

References: Stanford CS224N, Lecture 9: <https://web.stanford.edu/class/cs224n/slides/cs224n-2024-lecture09-pretraining.pdf>  
 Princeton COS 597G, Lecture 2: <https://www.cs.princeton.edu/courses/archive/fall22/cos597G/lectures/lec02.pdf>



**Sentence-Level Task**

- Sentence pair classification tasks:

**MNLI**      Premise: A soccer game with multiple males playing.  
 Hypothesis: Some men are playing a sport.      {entailment, contradiction, neutral}

**Multi-genre Natural Language Inference**: Predict the relationship between two sentences.

**QQP**      Q1: Where can I learn to invest in stocks?  
 Q2: How can I learn more about stocks?      {duplicate, not duplicate}

**Quora Question Pairs**: Detect paraphrase questions.

- Single sentence classification tasks:

**SST2**      rich veins of funny stuff in this movie      {positive, negative}

**Sentiment Analysis**

17

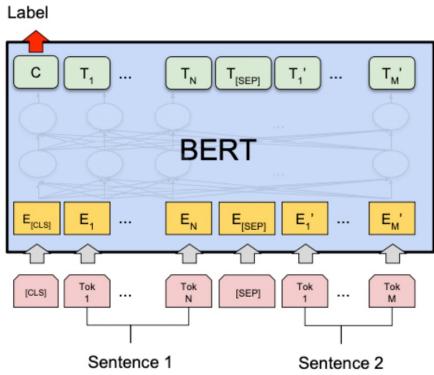
17

## Pretrain Once, Finetune Many Times

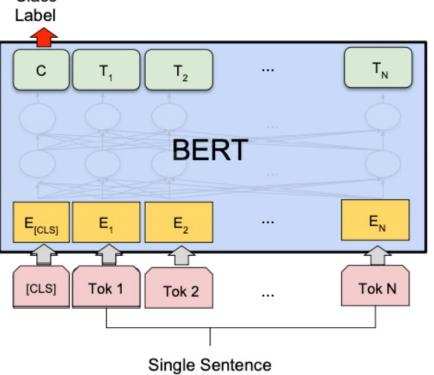
References: Stanford CS224N, Lecture 9: <https://web.stanford.edu/class/cs224n/slides/cs224n-2024-lecture09-pretraining.pdf>  
 Princeton COS 597G, Lecture 2: <https://www.cs.princeton.edu/courses/archive/fall22/cos597G/lectures/lec02.pdf>



**sentence-level tasks**



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



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

18

18

## Pretrain Once, Finetune Many Times

References: Stanford CS224N, Lecture 9: <https://web.stanford.edu/class/cs224n/slides/cs224n-2024-lecture09-pretraining.pdf>  
 Princeton COS 597G, Lecture 2: <https://www.cs.princeton.edu/courses/archive/fall22/cos597G/lectures/lec02.pdf>



### Token-Level Task

- Extractive question answering e.g., SQuAD (Rajpurkar et al., 2016)  
**Standard Question Answer Dataset:** Predict the answer to the question.

SQuAD

**Question:** The New York Giants and the New York Jets play at which stadium in NYC ?

**Context:** The city is represented in the National Football League by the New York Giants and the New York Jets , although both teams play their home games at **MetLife Stadium** in nearby East Rutherford , New Jersey , which hosted Super Bowl XLVIII in 2014 .

(Training example 29,883)

MetLife Stadium

- Named entity recognition (Tjong Kim Sang and De Meulder, 2003)  
**Named Entity Recognition:** Recognize the entity of each word.

CoNLL 2003 NER

John Smith lives in New York

B-PER I-PER O O B-LOC I-LOC

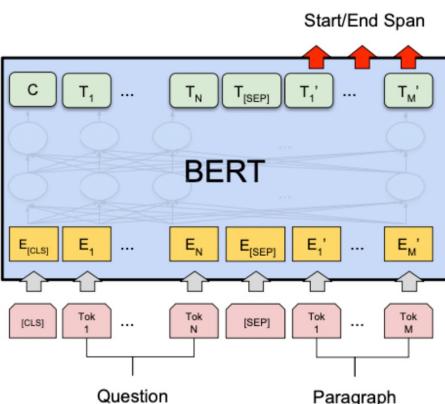
19

## Pretrain Once, Finetune Many Times

References: Stanford CS224N, Lecture 9: <https://web.stanford.edu/class/cs224n/slides/cs224n-2024-lecture09-pretraining.pdf>  
 Princeton COS 597G, Lecture 2: <https://www.cs.princeton.edu/courses/archive/fall22/cos597G/lectures/lec02.pdf>



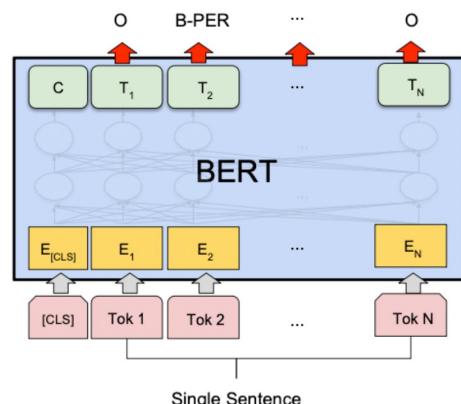
### token-level tasks



Start/End Span

BERT

Question      Paragraph



O      B-PER      ...      O

BERT

Single Sentence

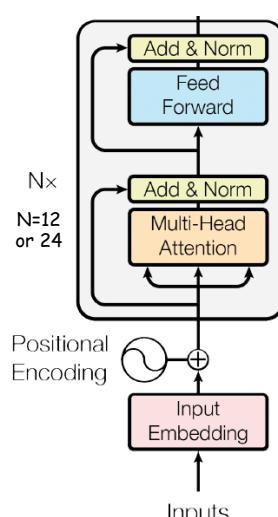
(c) Question Answering Tasks:  
SQuAD v1.1

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

20

## BERT was the State-of-The-Art

References: Stanford CS224N, Lecture 9: <https://web.stanford.edu/class/cs224n/slides/cs224n-2024-lecture09-pretraining.pdf>  
Princeton COS 597G, Lecture 2: <https://www.cs.princeton.edu/courses/archive/fall22/cos597G/lectures/lec02.pdf>



- BERT-base: 12 layers, 768-dim hidden size, 12 attention heads, 110M parameters
- BERT-large: 24 layers, 1024-dim hidden size, 16 attention heads, 340M parameters

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	Average
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
BERT <sub>BASE</sub>	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
BERT <sub>LARGE</sub>	<b>86.7/85.9</b>	<b>72.1</b>	<b>92.7</b>	<b>94.9</b>	<b>60.5</b>	<b>86.5</b>	<b>89.3</b>	<b>70.1</b>	<b>82.1</b>

- **Key issue with encoders:** Not a language model, i.e., does not naturally lead to autoregressive generation methods.

21

21

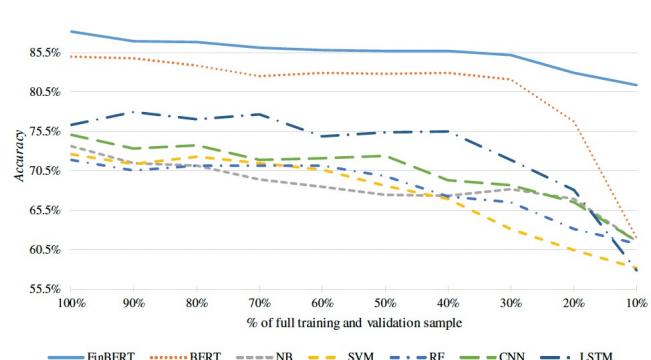
## FinBERT



- Pretrain BERT-base using financial datasets (4.9B tokens in total) with 4 P100 GPUs (100G memory):
  - Corporate annual and quarterly filings from SEC's EDGAR website (1994-2019).
  - Financial analyst reports from Thomson Intestext database (2003-2012).
  - Earnings conference call transcripts from the SeekingAlpha website (2004-2019).
- Finetuning and evaluation:
  - Sentiment analysis 10,000 sentences
  - 36% positive
  - 46% neutral
  - 18% negative
- Can FinBERT beat GPT-4 or Claude-3 in tasks related to financial texts?
  - How can we make fair comparisons?

**FinBERT: A large language model for extracting information from financial text**  
AH Huang, H Wang, Y Yang - Contemporary Accounting ..., 2023 - Wiley Online Library  
... model that adapts to the **finance** domain. We show that FinBERT incorporates **finance** knowledge and can better summarize contextual **information** in **financial texts**. Using a sample of ...  
☆ Save ⌂ Cite Cited by 144 Related articles Web of Science: 22 ☰

**Figure 1** Sentiment classification accuracy across sample sizes



% of full training and validation sample	FinBERT	BERT	NB	SVM	RF	CNN	LSTM
100%	~85.5%	~85.5%	~75.5%	~75.5%	~75.5%	~75.5%	~75.5%
90%	~85.0%	~85.0%	~75.0%	~75.0%	~75.0%	~75.0%	~75.0%
80%	~84.5%	~84.5%	~74.5%	~74.5%	~74.5%	~74.5%	~74.5%
70%	~84.0%	~84.0%	~74.0%	~74.0%	~74.0%	~74.0%	~74.0%
60%	~83.5%	~83.5%	~73.5%	~73.5%	~73.5%	~73.5%	~73.5%
50%	~83.0%	~83.0%	~73.0%	~73.0%	~73.0%	~73.0%	~73.0%
40%	~82.5%	~82.5%	~72.5%	~72.5%	~72.5%	~72.5%	~72.5%
30%	~82.0%	~82.0%	~72.0%	~72.0%	~72.0%	~72.0%	~72.0%
20%	~81.5%	~81.5%	~71.5%	~71.5%	~71.5%	~71.5%	~71.5%
10%	~81.0%	~81.0%	~71.0%	~71.0%	~71.0%	~71.0%	~71.0%

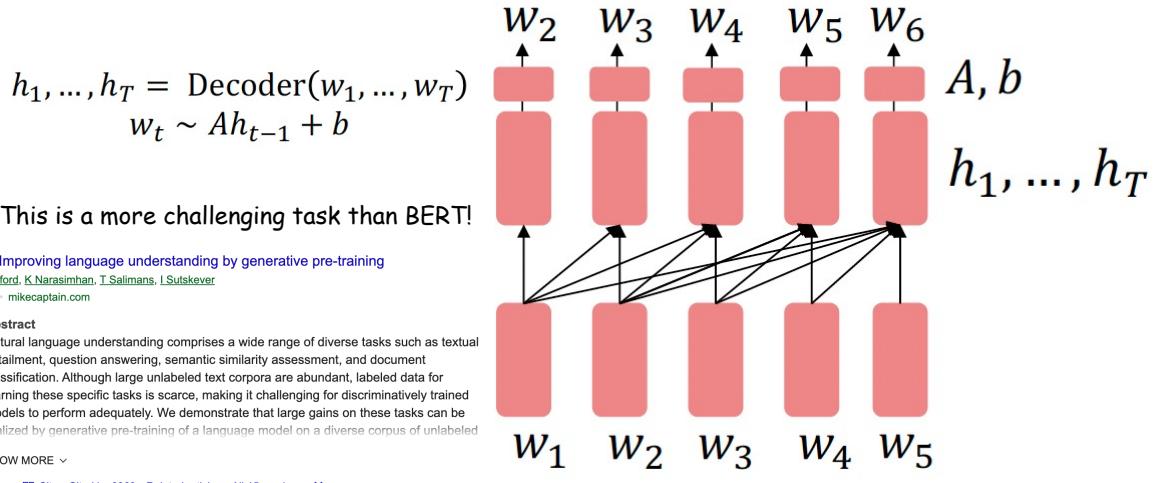
22

22

## Pretraining Decoders

References: Stanford CS224N, Lecture 9: <https://web.stanford.edu/class/cs224n/slides/cs224n-2024-lecture09-pretraining.pdf>

- Key idea: Pretrain decoders as language models  $\Pr(W_n | W_1, W_2, \dots, W_{n-1})$  via autoregression.



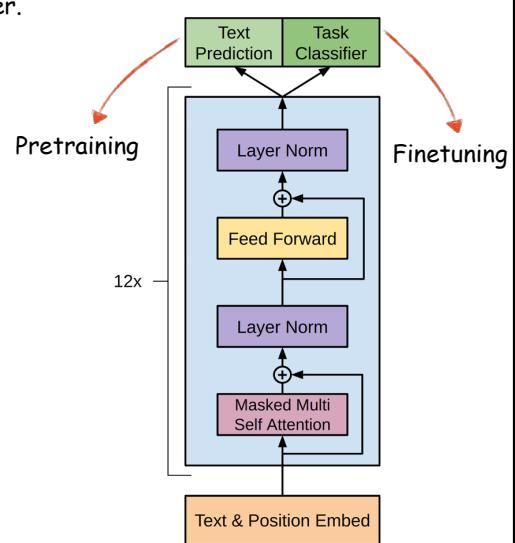
23

23

## GPT-1

References: Stanford CS224N, Lecture 9: <https://web.stanford.edu/class/cs224n/slides/cs224n-2024-lecture09-pretraining.pdf>

- Architecture: Only masked self-attention, but deeper and larger.
- 12 layers of transformer decoders, 117M parameters.
- 768-dim hidden states, 3072-dim MLP hidden layers.
- Trained on BooksCorpus of over 7,000 unique books.



[PDF] Improving language understanding by generative pre-training

A Radford, K Narasimhan, T Salimans, I Sutskever

2018 - mikedcaption.com

Abstract

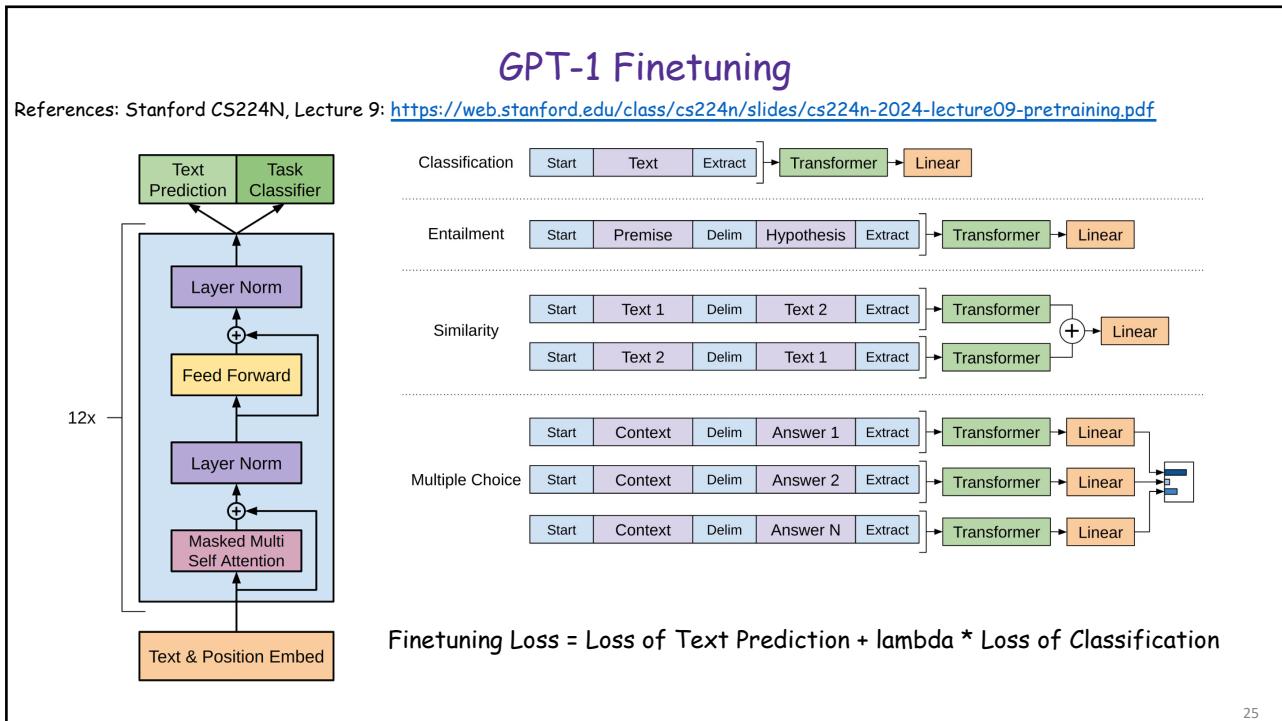
Natural language understanding comprises a wide range of diverse tasks such as textual entailment, question answering, semantic similarity assessment, and document classification. Although large unlabeled text corpora are abundant, labeled data for learning these specific tasks is scarce, making it challenging for discriminatively trained models to perform adequately. We demonstrate that large gains on these tasks can be realized by generative pre-training of a language model on a diverse corpus of unlabeled

SHOW MORE ▾

☆ Save 59 Cite Cited by 8363 Related articles All 15 versions ⓘ

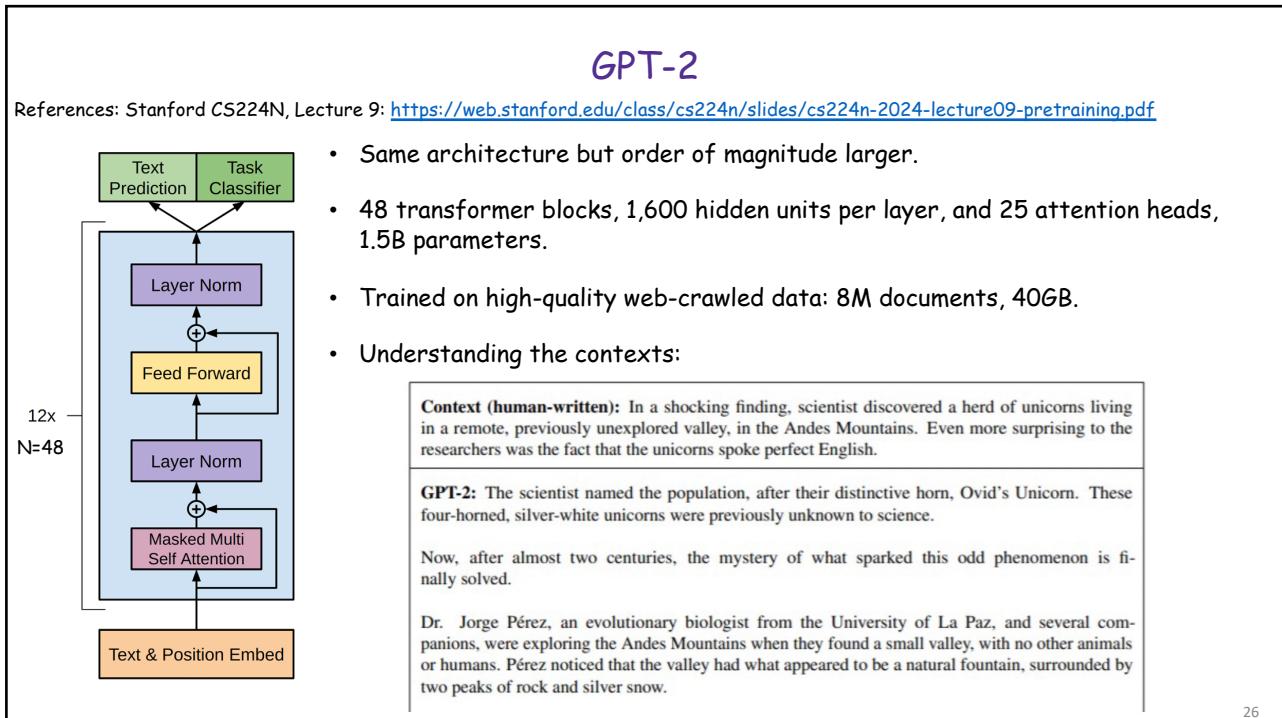
24

24



25

25

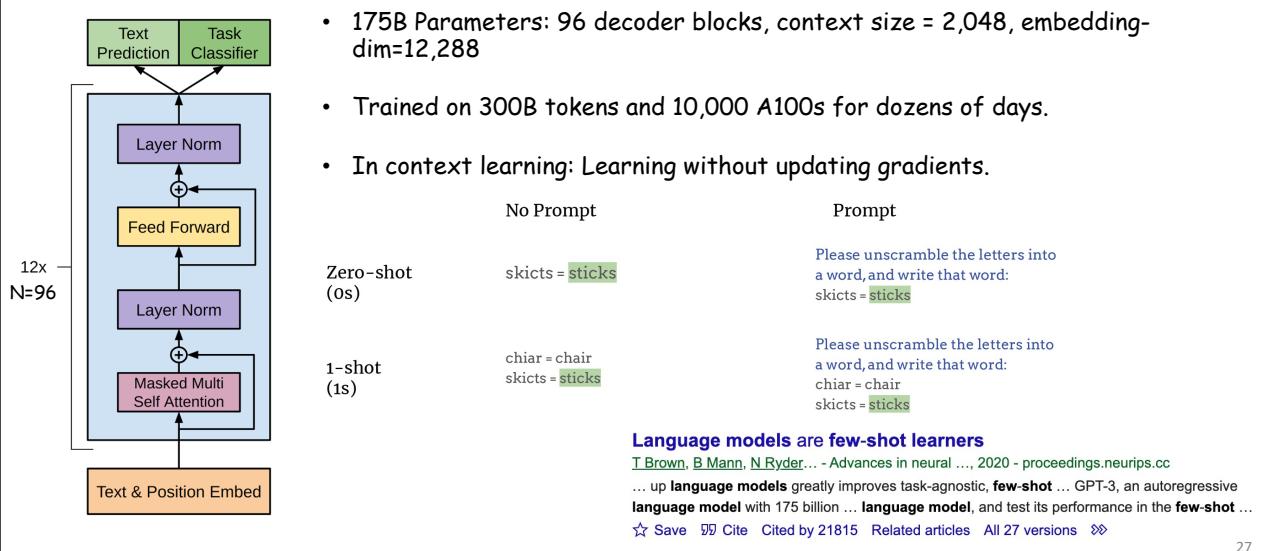


26

26

## GPT-3

References: Stanford CS224N, Lecture 9: <https://web.stanford.edu/class/cs224n/slides/cs224n-2024-lecture09-pretraining.pdf>  
 Princeton COS 597G, Lecture 4: <https://www.cs.princeton.edu/courses/archive/fall22/cos597G/lectures/lec04.pdf>



27

27

## Voice of Monetary Policy

### The Voice of Monetary Policy<sup>†</sup>

By YURIY GORODNICHENKO, THO PHAM, AND OLEKSANDR TALAVERA<sup>‡</sup>

We develop a deep learning model to detect emotions embedded in press conferences after the Federal Open Market Committee meetings and examine the influence of the detected emotions on financial markets. We find that, after controlling for the Federal Reserve's actions and the sentiment in policy texts, a positive tone in the voices of Federal Reserve chairs leads to significant increases in share prices. Other financial variables also respond to vocal cues from the chairs. Hence, how policy messages are communicated can move the financial market. Our results provide implications for improving the effectiveness of central bank communications. (JEL D83, E31, E44, E52, E58, F31, G14)

How can a president not be an actor?

—Ronald Reagan (1980)

As Chairman, I hope to foster a public conversation about what the Fed is doing to support a strong and resilient economy. And one practical step in doing so is to have a press conference like this after every one of our scheduled FOMC meetings. ... [This] is only about improving communications.  
 —Jerome Powell (2018)<sup>§</sup>

Monetary policy is 98 percent talk and 2 percent action, and communication is a big part.  
 —Ben Bernanke (2022)<sup>¶</sup>

- Use an MLP of 3 hidden layers to predict the voice tone of FOMC press conferences.

$$\text{VoiceTone} = \frac{\text{Positive answers} - \text{Negative answers}}{\text{Positive answers} + \text{Negative answers}},$$

- Use BERT to predict the sentiment of FOMC texts.

$$\text{TextSentiment} = \frac{\text{Dovish text} - \text{Hawkish text}}{\text{Dovish text} + \text{Hawkish text}},$$

- A positive tone of FR chairs leads to significant increases in share prices: How to say is as important as what to say.
- ?Seemed to suggest that using FinBERT saves the finetuning in sentiment analysis?

### The voice of monetary policy

[Y Gorodnichenko, T Pham, O Talavera - American Economic Review, 2023 - aeaweb.org](#)

... on recent advances in voice recognition technology and classify the voice tone of the Fed chairs into a spectrum of emotions. We, then, study how variations in voice tone (emotions) can ...

☆ Save 芻 Cite Cited by 118 Related articles All 30 versions Web of Science: 9 »

28

28

# Remote Work

## Remote Work across Jobs, Companies, and Space

Stephen Hansen, Peter John Lambert, Nicholas Bloom,  
Steven J. Davis, Raffaella Sadun & Bledi Taska

WORKING PAPER 31007 DOI 10.3386/w31007 ISSUE DATE March 2023

The pandemic catalyzed an enduring shift to remote work. To measure and characterize this shift, we examine more than 250 million job vacancy postings across five English-speaking countries. Our measurements rely on a state-of-the-art language-processing framework that we fit, test, and refine using 30,000 human classifications. We achieve 99% accuracy in flagging job postings that advertise hybrid or fully remote work, greatly outperforming dictionary methods and also outperforming other machine-learning methods. From 2019 to early 2023, the share of postings that say new employees can work remotely one or more days per week rose more than three-fold in the U.S. and by a factor of five or more in Australia, Canada, New Zealand and the U.K. These improvements are highly non-uniform across and within cities, industries, occupations, and companies. Even when zooming in on employers in the same industry competing for talent in the same occupations, we find large differences in the share of job postings that explicitly offer remote work.

- Pre-trained transformers are used for some **downstream tasks** (similarity measurement, concept detection, conception relationship characterization, text-metadata association, etc.).

- Use **DistilBERT** pre-trained on 1M text chunks of job vacancy postings to measure the **Work-from-homeness** of the 250 M jobs (**Work from Home Algorithmic Measure**), achieving 99% accuracy that outperforms dictionary-based methods.
- The number of WFM jobs has risen significantly since 2019 and it differs w.r.t. different industries.

### Remote work across jobs, companies, and space

[S Hansen](#), [PJ Lambert](#), [N Bloom](#), [SJ Davis](#), [R Sadun](#)... - 2023 - nber.org

The pandemic catalyzed an enduring shift to remote work. To measure and characterize this shift, we examine more than 250 million job vacancy postings across five English-speaking countries. Our measurements rely on a state-of-the-art language-processing framework that we fit, test, and refine using 30,000 human classifications. We achieve 99% accuracy in flagging job postings that advertise hybrid or fully remote work, greatly outperforming dictionary methods and also outperforming other machine-learning methods. From 2019 to ...

[☆ Save](#) [PDF Cite](#) [Cited by 36](#) [Related articles](#) [All 20 versions](#) [万分](#)

29

29

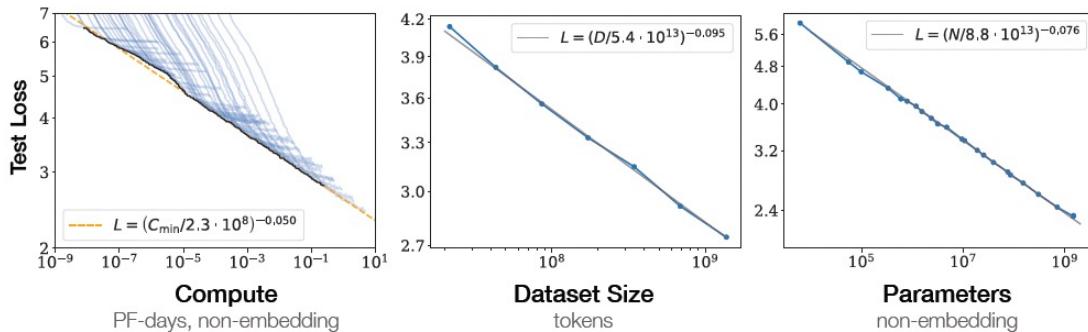
# Agenda

- Pre-training: BERT & GPT
- Large Language Models

30

30

## Revisit Scaling Laws



**Figure 1** Language modeling performance improves smoothly as we increase the model size, dataset size, and amount of compute used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

### Scaling laws for neural language models

J Kaplan, S McCandlish, T Henighan, T Brown... - arXiv preprint arXiv ..., 2020 - arxiv.org

... scaling laws for language model performance on the cross-entropy loss. The loss scales as a power-law with model ... Simple equations govern the dependence of overfitting on model/...

☆ Save 59 Cite Cited by 965 Related articles All 3 versions ☰

### Training compute-optimal large language models

J Hoffmann, S Borgeaud, A Mensch... - arXiv preprint arXiv ..., 2022 - arxiv.org

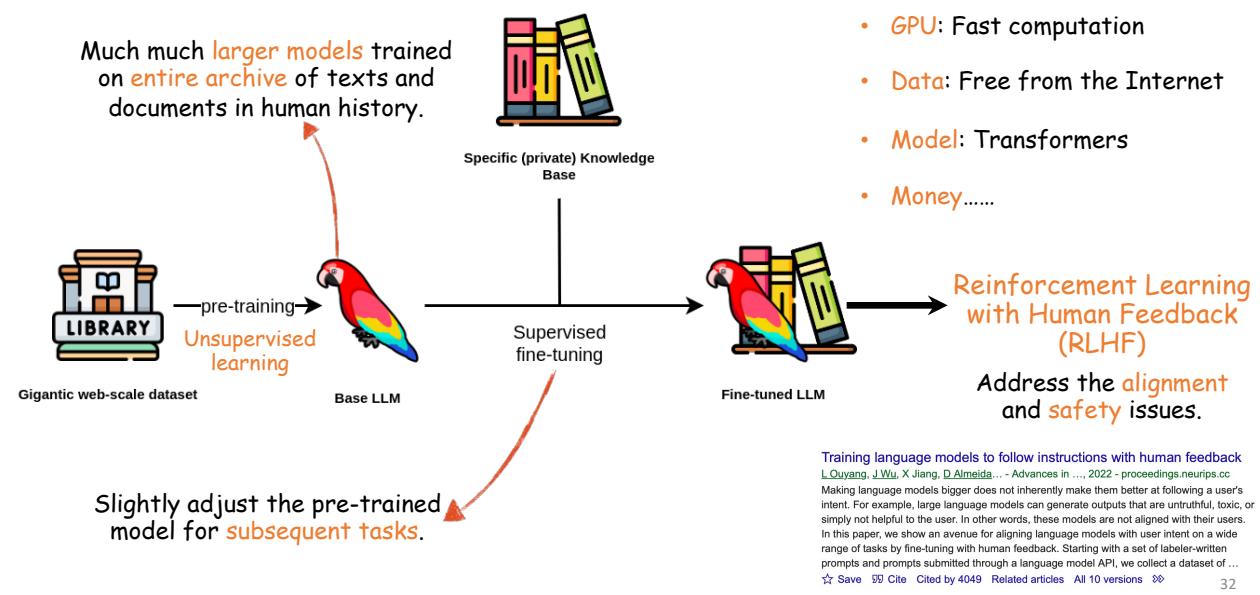
... We find that current large language models are significantly undertrained, a consequence of the recent focus on scaling language models whilst keeping the amount of training data ...

☆ Save 59 Cite Cited by 852 Related articles All 6 versions ☰

31

31

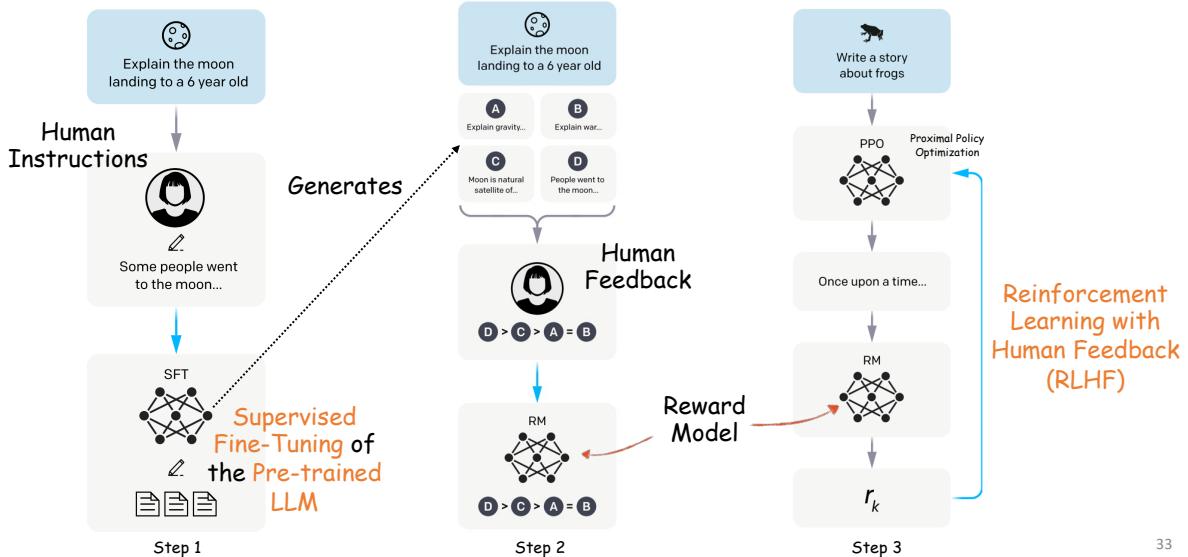
## Large Language Models (LLMs)



32

## SFT and RLHF

- We use **human feedback** (i.e., human playing with the language models) to improve it.



33

33

## Multi-Task Instruction Finetuning

References: Stanford CS224N, Lecture 10: <https://web.stanford.edu/class/cs224n/slides/cs224n-2024-lecture10-instruction-tuning-rlhf.pdf>

### Instruction finetuning

Please answer the following question.  
What is the boiling point of Nitrogen?

### Chain-of-thought finetuning

Answer the following question by reasoning step-by-step.  
The cafeteria had 23 apples. If they used 20 for lunch and bought 6 more, how many apples do they have?

### Multi-task instruction finetuning (1.8K tasks)

#### Inference: generalization to unseen tasks

Q: Can Geoffrey Hinton have a conversation with George Washington?  
Give the rationale before answering.

Language model

-320.4F

The cafeteria had 23 apples originally. They used 20 to make lunch. So they had  $23 - 20 = 3$ . They bought 6 more apples, so they have  $3 + 6 = 9$ .

Geoffrey Hinton is a British-Canadian computer scientist born in 1947. George Washington died in 1799. Thus, they could not have had a conversation together. So the answer is "no".

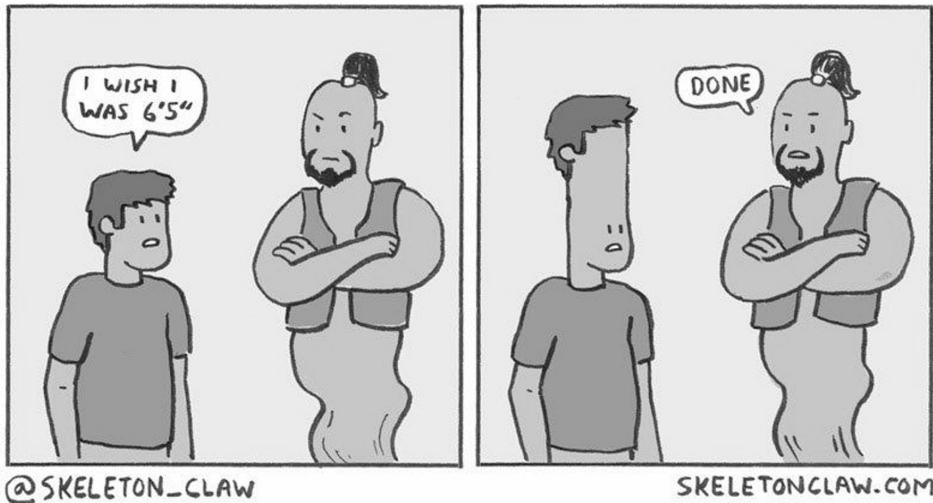
34

34

## AI Misalignment

- **Misalignment:** AI behaves in a way humans do not want.

GENIE 2



35

35

## Reward Model

References: Stanford CS224N, Lecture 10: <https://web.stanford.edu/class/cs224n/slides/cs224n-2024-lecture10-instruction-tuning-rlhf.pdf>

- For a language model output sample  $s$  on a specific task (e.g., summarization), try to obtain a human reward  $R(s)$ , the higher the better.

SAN FRANCISCO,  
California (CNN) --  
A magnitude 4.2  
earthquake shook the  
San Francisco  
...  
overtake unstable  
objects.

An earthquake hit  
San Francisco.  
There was minor  
property damage,  
but no injuries.

$$s_1 \\ R(s_1) = 8.0$$

The Bay Area has  
good weather but is  
prone to  
earthquakes and  
wildfires.

$$s_2 \\ R(s_2) = 1.2$$

- Ask the labellers to make pairwise comparisons, train a reward model and try to maximize:

$$\mathbb{E}_{\hat{s} \sim p_\theta(s)} [R(\hat{s})]$$

36

36

## Instruct-GPT

PROMPT *Explain the moon landing to a 6 year old in a few sentences.*

COMPLETION GPT-3

Explain the theory of gravity to a 6 year old.

Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.

InstructGPT

People went to the moon, and they took pictures of what they saw, and sent them back to the earth so we could all see them.

37

37

## Instruct-GPT

PROMPT *Write a short poem about a wise frog.*

COMPLETION GPT-3

Write a short story in which a character has two different names.

Write a short story in which you try to get something back that you have lost.

Write a short story in which a character has a bad dream.

InstructGPT

The frog is a symbol of wisdom  
 He knows all the secrets of the world  
 He is a master of disguise  
 And a great teacher of life  
 He is a symbol of transformation  
 And the bringer of change  
 He is the frog who has seen it all  
 And knows the meaning of it all

38

38

## Value of Fine-Tuning and RLHF

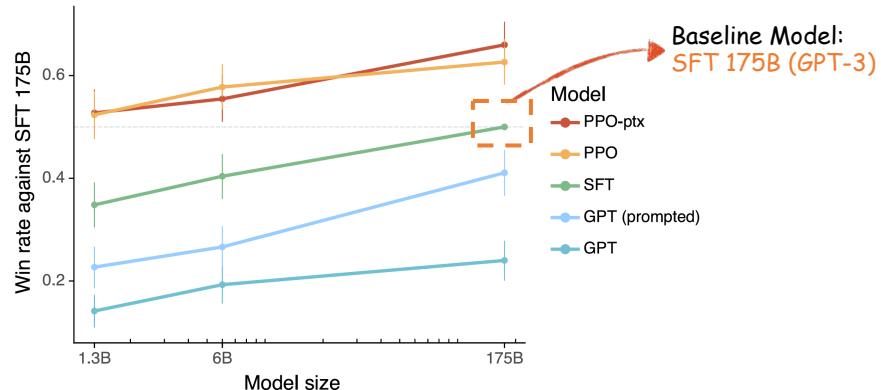


Table 6: Dataset sizes, in terms of number of prompts.

SFT Data			RM Data			PPO Data		
split	source	size	split	source	size	split	source	size
train	labeler	11,295	train	labeler	6,623	train	customer	31,144
train	customer	1,430	train	customer	26,584	valid	customer	16,185
valid	labeler	1,550	valid	labeler	3,488			
valid	customer	103	valid	customer	14,399			

39

39

## Issues with RLHF

References: Stanford CS224N, Lecture 10: <https://web.stanford.edu/class/cs224n/slides/cs224n-2024-lecture10-instruction-tuning-rlhf.pdf>

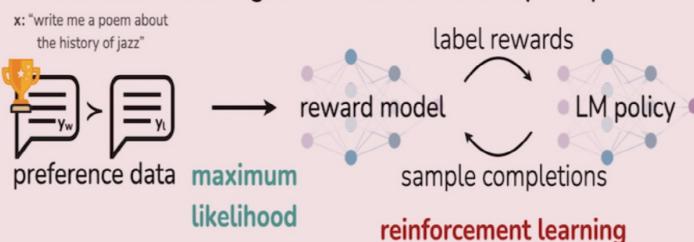
- Human preferences are unreliable, so the LLM are rewarded to produce responses that seem authoritative and helpful, regardless of truth: Make-up facts and hallucinations.
- Reinforcement learning is unstable.
  - Direct preference optimization (DPO).
  - Open-source LLMs now mostly use DPO.

TECHNOLOGY

Google shares drop \$100 billion after its new AI chatbot makes a mistake

February 9, 2023 · 10:15 AM ET

### Reinforcement Learning from Human Feedback (RLHF)



### Direct Preference Optimization (DPO)



40

40

## Hallucination

**The ChatGPT Lawyer Explains Himself**

In a cringe-inducing court hearing, a lawyer who relied on A.I. to craft a motion full of made-up case law said he "did not comprehend" that the chat bot could lead him astray.

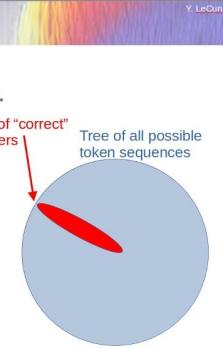


Share full article 267

Steven A. Schwartz told a judge considering sanctions that the episode had been "deeply embarrassing." Jefferson Siegel for The New York Times

**Unpopular Opinion about AR-LLMs**

- ▶ Auto-Regressive LLMs are **doomed**.
- ▶ They cannot be made factual, non-toxic, etc.
- ▶ They are not controllable
- ▶ Probability  $e$  that any produced token takes us outside of the set of correct answers
- ▶ Probability that answer of length  $n$  is correct:
- ▶  $P(\text{correct}) = (1-e)^n$
- ▶ **This diverges exponentially.**
- ▶ **It's not fixable (without a major redesign).**



Tree of "correct" answers  
Tree of all possible token sequences

You Tell me about AeroGlide UltraSlim Smart Toothbrush by Boie

ChatGPT The AeroGlide UltraSlim Smart Toothbrush by Boie is a state-of-the-art oral care device designed to provide an exceptional cleaning experience. It incorporates advanced sonic technology for a thorough clean, featuring a slim and sleek design that ensures ease of use and maneuverability. This toothbrush comes equipped with various smart features aimed at optimizing your brushing routine, enhancing the effectiveness and efficiency of your oral hygiene practices.

Boie really exists but not the product.

41

## Emergent Abilities

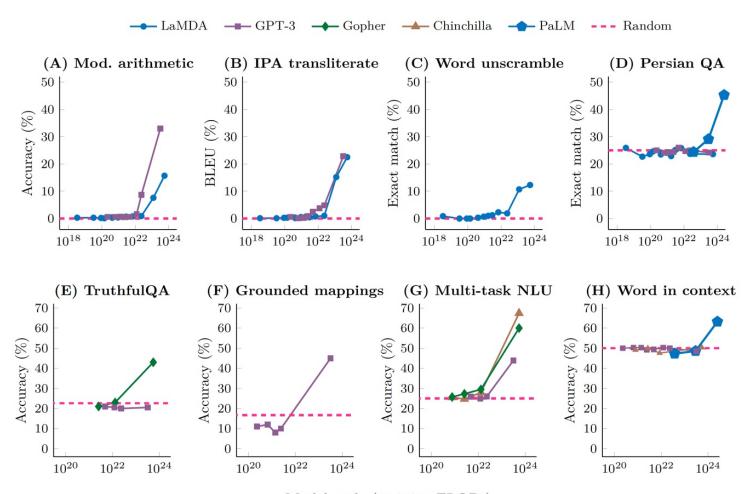
- **Emergent Abilities:** An ability not present in smaller models but present in larger models.
- **Phase-change in physics:** Quantitative changes in the system result in qualitative changes in behavior.

**Emergent abilities of large language models**

J.Wei, Y.Tay, R.Bommasani, C.Raffel, B.Zoph... - arXiv preprint arXiv ..., 2022 - a

... an **ability** to be **emergent** if it is not present in smaller **models** but is present in larger **models**. ... We have discussed **emergent abilities of language models**, for which meaningful ...

Save Cite Cited by 1256 Related articles All 11 versions



Task	Model	10 <sup>18</sup> FLOPs	10 <sup>20</sup> FLOPs	10 <sup>22</sup> FLOPs	10 <sup>24</sup> FLOPs
(A) Mod. arithmetic	LaMDA	~1%	~1%	~1%	~15%
	GPT-3	~1%	~1%	~10%	~35%
	Gopher	~1%	~1%	~1%	~15%
	Chinchilla	~1%	~1%	~1%	~15%
	PaLM	~1%	~1%	~1%	~15%
	Random	~1%	~1%	~1%	~1%
(B) IPA transliterate	LaMDA	~1%	~1%	~1%	~15%
	GPT-3	~1%	~1%	~1%	~15%
	Gopher	~1%	~1%	~1%	~15%
	Chinchilla	~1%	~1%	~1%	~15%
	PaLM	~1%	~1%	~1%	~15%
	Random	~1%	~1%	~1%	~1%
(C) Word unscramble	LaMDA	~1%	~1%	~1%	~15%
	GPT-3	~1%	~1%	~1%	~15%
	Gopher	~1%	~1%	~1%	~15%
	Chinchilla	~1%	~1%	~1%	~15%
	PaLM	~1%	~1%	~1%	~15%
	Random	~1%	~1%	~1%	~1%
(D) Persian QA	LaMDA	~15%	~15%	~15%	~45%
	GPT-3	~15%	~15%	~15%	~45%
	Gopher	~15%	~15%	~15%	~45%
	Chinchilla	~15%	~15%	~15%	~45%
	PaLM	~15%	~15%	~15%	~45%
	Random	~15%	~15%	~15%	~45%
(E) TruthfulQA	LaMDA	~15%	~15%	~15%	~45%
	GPT-3	~15%	~15%	~15%	~45%
	Gopher	~15%	~15%	~15%	~45%
	Chinchilla	~15%	~15%	~15%	~45%
	PaLM	~15%	~15%	~15%	~45%
	Random	~15%	~15%	~15%	~45%
(F) Grounded mappings	LaMDA	~15%	~15%	~15%	~45%
	GPT-3	~15%	~15%	~15%	~45%
	Gopher	~15%	~15%	~15%	~45%
	Chinchilla	~15%	~15%	~15%	~45%
	PaLM	~15%	~15%	~15%	~45%
	Random	~15%	~15%	~15%	~45%
(G) Multi-task NLU	LaMDA	~15%	~15%	~15%	~45%
	GPT-3	~15%	~15%	~15%	~45%
	Gopher	~15%	~15%	~15%	~45%
	Chinchilla	~15%	~15%	~15%	~45%
	PaLM	~15%	~15%	~15%	~45%
	Random	~15%	~15%	~15%	~45%
(H) Word in context	LaMDA	~15%	~15%	~15%	~45%
	GPT-3	~15%	~15%	~15%	~45%
	Gopher	~15%	~15%	~15%	~45%
	Chinchilla	~15%	~15%	~15%	~45%
	PaLM	~15%	~15%	~15%	~45%
	Random	~15%	~15%	~15%	~45%

42

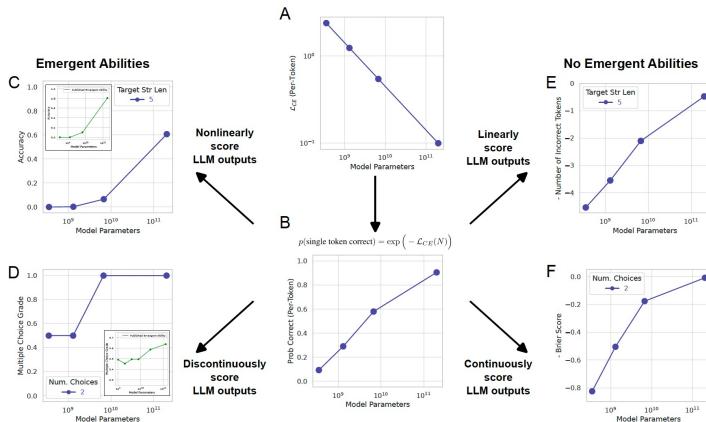
42

## Are Emergent Abilities of LLMs a Mirage?

$$\text{Multiple Choice Grade} \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if highest probability mass on correct option} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Exact String Match} \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if output string exactly matches target string} \\ 0 & \text{otherwise} \end{cases}$$

- **Emergent Abilities** may be attributed to the choice of nonlinear or discontinuous metrics, whereas linear or continuous metrics produce smooth performance changes.



NeurIPS 2023 Outstanding Main Track Paper

### Power-law of Scaling

$$\mathcal{L}_{CE}(N) = \left(\frac{N}{c}\right)^{\alpha}$$

$$p(\text{single token correct}) = \exp(-\mathcal{L}_{CE}(N)) = \exp\left(-(N/c)^{\alpha}\right)$$

$$\text{Accuracy}(N) \approx p_N(\text{single token correct})^{\text{num. of tokens}} = \exp\left(-(N/c)^{\alpha}\right)^L$$

Figure C

Figure E

$$\text{Token Edit Distance}(N) \approx L \left(1 - p_N(\text{single token correct})\right) = L \left(1 - \exp\left(-(N/c)^{\alpha}\right)\right)$$

Are emergent abilities of large language models a mirage?

R Schaeffer, B Miranda... - Advances in Neural ..., 2024 - proceedings.neurips.cc  
... be interpreted as claiming that large language models cannot display emergent abilities;  
rather, our message is that some previously claimed emergent abilities appear to be mirages ...

☆ Save 99 Cite Cited by 129 Related articles All 9 versions

43

43

## In-Context Learning

- The larger the model, the more context you need to provide to respond to the question correctly.
- Asking a 5-year-old  $1 + 1 = ?$  vs. Asking a mathematician  $1 + 1 = ?$

No Prompt

Zero-shot  
(0s)

skicts = sticks

1-shot  
(1s)

chiar = chair  
skicts = sticks

Few-shot  
(FS)

chiar = chair  
[...]  
pciinc = picnic  
skicts = sticks

Prompt

Please unscramble the letters into a word, and write that word:  
skicts = sticks

Please unscramble the letters into a word, and write that word:  
chiar = chair  
skicts = sticks

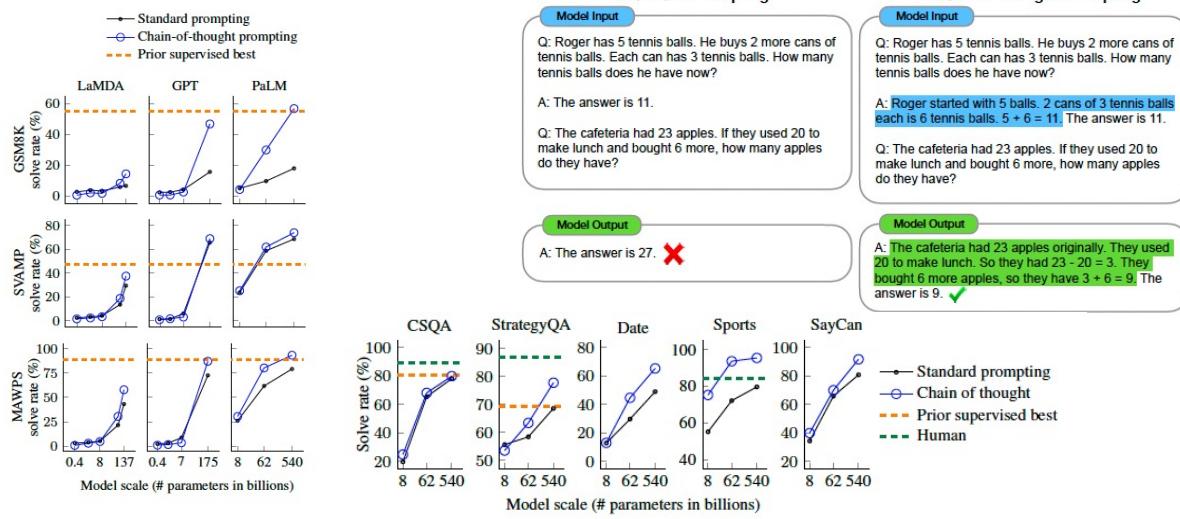
Please unscramble the letters into a word, and write that word:  
chiar = chair  
[...]  
pciinc = picnic  
skicts = sticks

44

44

## Chain-of-Thought (CoT)

- A series of intermediate reasoning steps significantly improves the ability of large language models for complex reasoning.



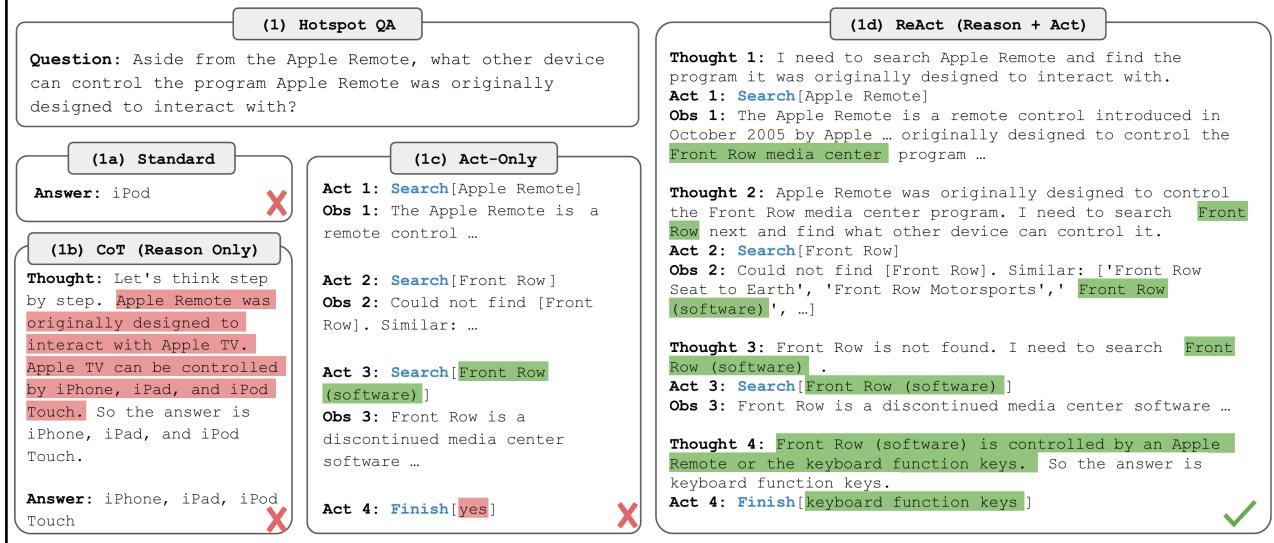
Reference (CoT Paper): Chain-of-Thought Prompting Elicits Reasoning in Large Language Models (NeurIPS 2022)

45

45

## ReAct

- ReAct = Reason (CoT) + Act (Obtain external information)



Reference (ReAct Paper): ReAct: Synergizing Reasoning and Acting in Language Models (ICLR 2023)

46

46

## Reflexion: Self-Reflecting LLM

- Reflexion = ReAct + Reinforce language agents not by updating weights, but through linguistic feedback

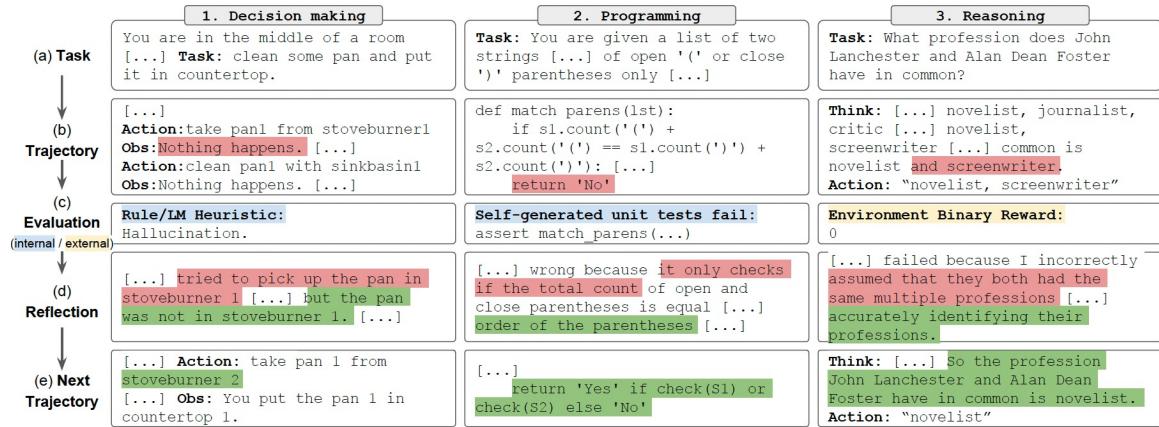
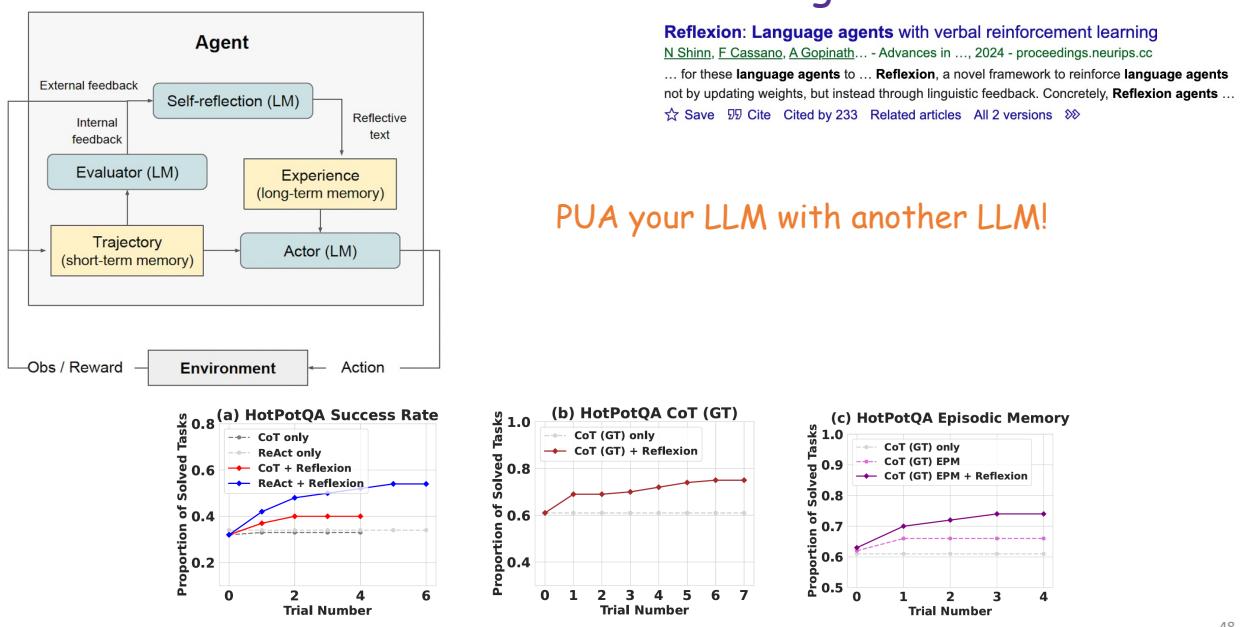


Figure 1: Reflexion works on decision-making [4.1], programming [4.3], and reasoning [4.2] tasks.

Reference (Reflexion Paper): Reflexion: Language Agents with Verbal Reinforcement Learning (NeurIPS 2023)

47

## Reflexion: Self-Reflecting LLM



48

48

## Generative AI for Econ/Social Science Researchers

*Journal of Economic Literature* 2023, 61(4), 1281–1317  
<https://doi.org/10.1257/jel.20231736>

### Generative AI for Economic Research: Use Cases and Implications for Economists<sup>1</sup>

ANTON KORINEK<sup>2</sup>

*Generative artificial intelligence (AI) has the potential to revolutionize research. I analyze how large language models (LLMs) such as ChatGPT can assist economists by describing dozens of use cases in six areas: ideation and feedback, writing, background research, data analysis, coding, and mathematical derivations. I provide general instructions and demonstrate specific examples of how to take advantage of each of these, classifying the LLM capabilities from experimental to highly useful. I argue that economists can reap significant productivity gains by taking advantage of generative AI to automate micro-tasks. Moreover, these gains will grow as the performance of AI systems continues to improve. I also speculate on the longer-term implications of AI-powered cognitive automation for economic research. The online resources associated with this paper explain how to get started and will provide regular updates on the latest capabilities of generative AI in economics. (JEL A11, C45, D83, I23, O33)*

**Impact of a CS Paper = Problem Importance \* Technical Novelty \* Performance Improvement**

**Impact of a Business Paper = Problem Importance \* Identification Rigor \* Insight Novelty**

49

49

## LLM for Mathematical Discoveries

[HTML] Mathematical discoveries from program search with large language models  
 B.Romera-Paredes, M.Barekatian, A.Novikov, M.Balog, M.P.Kumar, E.Dupont, FJR Ruiz...  
 Nature, 2024 · nature.com

### Abstract

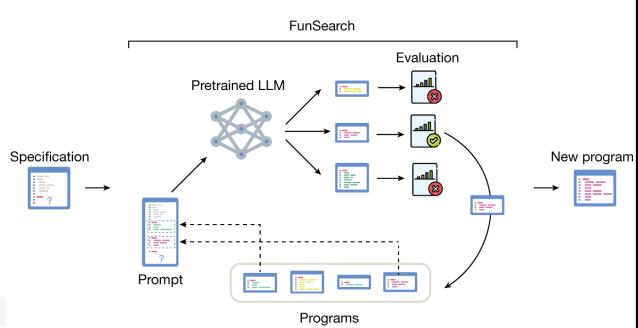
Large language models (LLMs) have demonstrated tremendous capabilities in solving complex tasks, from quantitative reasoning to understanding natural language. However, LLMs sometimes suffer from confabulations (or hallucinations), which can result in them making plausible but incorrect statements. This hinders the use of current large models in scientific discovery. Here we introduce FunSearch (short for searching in the function space), an evolutionary procedure based on pairing a pretrained LLM with a systematic evaluator. We demonstrate the effectiveness of this approach to surpass the best-known results in important problems, pushing the boundary of existing LLM-based approaches. Applying FunSearch to a central problem in extremal combinatorics—the cap set problem—we discover new constructions of large cap sets going beyond the best-known ones, both in finite dimensional and asymptotic cases. This shows that it is possible to make discoveries for established open problems using LLMs. We showcase the generality of FunSearch by applying it to an algorithmic problem, online bin packing, finding new heuristics that improve on widely used baselines. In contrast to most computer search approaches, FunSearch searches for programs that describe how to solve a problem, rather than what the solution is. Beyond being an effective and scalable strategy, discovered programs tend to be more interpretable than raw solutions, enabling feedback loops between domain experts and FunSearch, and the deployment of such programs in real-world applications.

nature.com

SHOW LESS ^

☆ Save 99 Cite Cited by 54 Related articles All 7 versions

- LLM helps find new solutions to challenging combinatorial problems that **surpass the best-known results**.
- An **evolutionary procedure** that pairs a **pretrained LLM** with a **systematic evaluator**.



50

50

Group	Occupations with highest exposure	% Exposure
Human $\alpha$	Interpreters and Translators	76.5
	Survey Researchers	75.0
	Poets, Lyricists and Creative Writers	68.8
	Animal Scientists	66.7
	Public Relations Specialists	66.7
Human $\beta$	Survey Researchers	84.4
	Writers and Authors	82.5
	Interpreters and Translators	82.4
	Public Relations Specialists	80.6
	Animal Scientists	77.8
Human $\zeta$	Mathematicians	100.0
	Tax Preparers	100.0
	Financial Quantitative Analysts	100.0
	Writers and Authors	100.0
	Web and Digital Interface Designers	100.0
	<i>Humans labeled 15 occupations as "fully exposed."</i>	
Model $\alpha$	Mathematicians	100.0
	Correspondence Clerks	95.2
	Blockchain Engineers	94.1
	Court Reporters and Simultaneous Captioners	92.9
	Proofreaders and Copy Markers	90.9
Model $\beta$	Mathematicians	100.0
	Blockchain Engineers	97.1
	Court Reporters and Simultaneous Captioners	96.4
	Proofreaders and Copy Markers	95.5
	Correspondence Clerks	95.2
Model $\zeta$	Accountants and Auditors	100.0
	News Analysts, Reporters, and Journalists	100.0
	Legal Secretaries and Administrative Assistants	100.0
	Clinical Data Managers	100.0
	Climate Change Policy Analysts	100.0
	<i>The model labeled 86 occupations as "fully exposed."</i>	
Highest variance	Search Marketing Strategists	14.5
	Graphic Designers	13.4
	Investment Fund Managers	13.0
	Financial Managers	13.0
	Insurance Appraisers, Auto Damage	12.6

Job Zone Required	Education Required	Example Occupations	Median Income	Tot Emp (000s)	H $\alpha$	M $\alpha$	H $\beta$	M $\beta$	H $\zeta$	M $\zeta$
1	None or little (0-3 months)	High school diploma or GED (optional)	\$30,230	13,100	0.03	0.04	0.06	0.06	0.09	0.08
2	Some (3-12 months)	High school diploma	\$38,215	73,962	0.07	0.12	0.16	0.20	0.24	0.27
3	Medium (1-2 years)	Vocational school, on-the-job training, or associate's degree	\$54,815	37,881	0.11	0.14	0.26	0.32	0.41	0.51
4	Considerable (2-4 years)	Bachelor's degree	\$77,345	56,833	0.23	0.18	0.47	0.51	0.71	0.85
5	Extensive (4+ years)	Master's degree or higher	\$81,980	21,221	0.23	0.13	0.43	0.45	0.63	0.76

**Jobs Exposed to Generative AI**

More "intelligent" jobs are more exposed to generative AI!

**Gpts are gpts: An early look at the labor market impact potential of large language models**

[T Eloundou, S Manning, P Mishkin, D Rock - arXiv preprint arXiv ..., 2023 - arxiv.org](#)

... potential implications of large language models (LLMs), such as Generative Pretrained Transformers (GPTs), on the US labor market, ... both human expertise and GPT-4 classifications. ...

[☆ Save](#) [99 Cite](#) [Cited by 410](#) [Related articles](#) [All 20 versions](#) [»»](#)

Reference: <https://arxiv.org/pdf/2303.10130.pdf>

51

**Generative AI Improves Productivity and Equality**

[RESEARCH ARTICLE](#) | CHATGPT

[f](#) [t](#) [in](#) [g](#) [o](#) [n](#)

**Experimental evidence on the productivity effects of generative artificial intelligence**

**SHAKED NOY AND WHITNEY ZHANG** [Authors Info & Affiliations](#)

**SCIENCE** • 13 Jul 2023 • Vol 381, issue 6654 • pp. 187-192 • DOI: 10.1126/science.adb2586

36,231 [1](#) [Bell](#) [Share](#) [Print](#)

**Editor's summary**

Automation has historically displaced human workers in factories (e.g., automotive manufacturing) or in performing routine computational tasks. Will generative artificial intelligence (AI) tools such as ChatGPT disrupt the labor market by making educated professionals obsolete, or will these tools complement their skills and enhance productivity? Noy and Zhang examined this issue in an experiment that recruited college-educated professionals to complete incentivized writing tasks. Participants assigned to use ChatGPT were more productive, efficient, and enjoyed the tasks more. Participants with weaker skills benefited the most from ChatGPT, which carries policy implications for efforts to reduce productivity inequality through AI. —EEU

**Experimental evidence on the productivity effects of generative artificial intelligence**

S.Noy, W.Zhang - Science, 2023 - science.org

... the productivity effects of a generative artificial intelligence (AI)... In a preregistered online experiment, we assigned ... Our results show that ChatGPT substantially raised productivity: The ...

[☆ Save](#) [99 Cite](#) [Cited by 262](#) [Related articles](#) [All 10 versions](#) [»»](#)

**PNAS** [BRIEF REPORT](#) | POLITICAL SCIENCES [OPEN ACCESS](#) [Check for updates](#)

**ChatGPT outperforms crowd workers for text-annotation tasks**

Fabrizio Gilardi\*, Meysam Alizadeh\*, and Mait Kubil†

Edited by Mary Waters, Harvard University, Cambridge, MA; received March 27, 2023; accepted June 2, 2023

Many NLP applications require manual text annotations for a variety of tasks, notably to train classifiers or evaluate the performance of unsupervised models. Depending on the size and degree of complexity, the tasks may be conducted by crowd workers on platforms such as Amazon Mechanical Turk (MTurk). We report on a recent experiment. Using four samples of tweets and news articles ( $n = 6,183$ ), we show that ChatGPT outperforms crowd workers for several annotation tasks, including relevance, stance, topics, and frame detection. Across the four datasets, the zero-shot accuracy of ChatGPT exceeds that of crowd workers by about 25 percentage points on average, while ChatGPT's intercoder agreement exceeds that of both crowd workers and trained annotators for all tasks. Moreover, the per-annotation cost of ChatGPT is less than \$0.003—about thirty times cheaper than MTurk. These results demonstrate the potential of large language models to drastically increase the efficiency of text classification.

ChatGPT | text classification | large language models | human annotations | text as data

- GenAI improves the productivity of knowledge workers.
- GenAI reduces the productivity inequality among the workers.

52

# Generative AI Powered Conversational Assistants

## Generative AI at Work

Erik Brynjolfsson, Danielle Li & Lindsey R. Raymond

WORKING PAPER 31161 DOI 10.3386/w31161 ISSUE DATE April 2023 REVISION DATE November 2023

New AI tools have the potential to change the way workers perform and learn, but little is known about their impacts on the job. In this paper, we study the staggered introduction of a generative AI-based conversational assistant using data from 5,179 customer support agents. Access to the tool increases productivity, as measured by issues resolved per hour, by 14% on average, including a 34% improvement for novice and low-skilled workers but with minimal impact on experienced and highly skilled workers. We provide suggestive evidence that the AI model disseminates the best practices of more able workers and helps newer workers move down the experience curve. In addition, we find that AI assistance improves customer sentiment, increases employee retention, and may lead to worker learning. Our results suggest that access to generative AI can increase productivity, with large heterogeneity in effects across workers.

**Generative AI at work**  
E Brynjolfsson, D Li, LR Raymond - 2023 - nber.org  
... In this paper, we study the adoption of a **generative AI** tool that provides conversational ...  
of **generative AI** when deployed at scale in the workplace. We find that access to **AI** assistance ...  
☆ Save ⚡ Cite Cited by 222 Related articles All 16 versions ☰

53

# Search Engine Optimization

MARKETING SCIENCE  
Vol. 41, No. 3, May–June 2022, pp. 441–462  
ISSN 0732-2899 (print), ISSN 1546-546X (online)

Frontiers: Supporting Content Marketing with Natural Language Generation  
Martin Reisenbichler,<sup>a,\*</sup> Thomas Reutterer,<sup>b,\*</sup> David A. Schweidel,<sup>c</sup> Daniel Dan<sup>c</sup>

<sup>a</sup>Department of Marketing, Vienna University of Economics and Business, Vienna A-1020, Austria; <sup>b</sup>Goizueta Business School, Marketing Area, Emory University, Atlanta, Georgia 30322; <sup>c</sup>School of Applied Data Science, Modul University, Vienna, Vienna A-1190, Austria  
(Corresponding authors)  
Contact: martin.reisenbichler@wuu.ac.at (MR); thomas.reutterer@vumc.at; <https://orcid.org/0000-0003-1276-8239> (TR); dschweidel@emory.edu, <https://orcid.org/0000-0003-2665-3272> (DD); daniel.dan@modul.ac.at (DD)

Received: June 30, 2021  
Revised: September 5, 2021  
Accepted: September 15, 2021  
Published Online in Articles in Advance: February 25, 2022  
<https://doi.org/10.1287/mksc.2022.1354>  
Copyright © 2022 INFORMS

**Abstract.** Advances in natural language generation (NLG) have facilitated technologies such as digital voice assistants and chatbots. In this research, we demonstrate how NLG can support content marketing by using it to draft content for the landing pages of a website in search engine optimization (SEO). Traditional SEO projects rely on hand-crafted content that is both time consuming and costly to produce. To address the costs associated with producing SEO content, we propose a technology that uses state-of-the-art NLG and demonstrate that the content-writing machine can generate unique, human-like SEO content. As part of our research, we demonstrate that although the machine-generated content is designed to perform well in search engines, the role of the human editor remains essential. Comparing the resulting content with human refinement to traditional human-written SEO texts, we find that the revised, machine-generated texts are virtually indistinguishable from those created by SEO experts along a number of human performance metrics. We also compare the cost of producing SEO content using our approach and show that the resulting SEO content outperforms that created by human writers (including SEO experts) in search engine rankings. Additionally, we illustrate how our approach can substantially reduce the production costs associated with content marketing, increasing their return on investment.

History K. Sudhir served as the senior editor and Olivier Touba served as associate editor for this article. This paper was accepted through the Marketing Science: Frontiers review process.  
Supplemental Material: Data and the web appendices are available at <https://doi.org/10.1287/mksc.2022.1354>.

Keywords: SEO • content marketing • natural language generation • transfer learning

**Frontiers: Supporting content marketing with natural language generation**  
M Reisenbichler, T Reutterer... - Marketing ..., 2022 - pubsonline.informs.org  
... can support **content marketing** by using it to draft **content** for the ... **content** that is both time consuming and costly to produce. To address the costs associated with producing SEO **content**, ...  
☆ Save ⚡ Cite Cited by 43 Related articles All 6 versions Web of Science: 8 ☰

**SEO Content Writing Machine (Automated)**

The diagram illustrates the SEO Content Writing Machine (Automated) process. It starts with a 'Human Select Keyword' box, which feeds into a 'Human Select and revise sel\_txt' box. The process involves several automated steps: 'Ranking and Links Crawling' leading to a 'Search engine' (represented by a circle with a dot), which then leads to 'Content scraping Websites' (represented by a circle with a dot and a list of 'top\_txt' items). This leads to 'GPT-2' (represented by a circle with a dot and a list of 'gen\_txt' items). 'GPT-2' then leads to 'Fine-tuning and Content Generation' (represented by a circle with a dot and a list of 'gen\_txt' items). Finally, it leads to 'Content Selection and Output Quality score' (represented by a circle with a dot and a list of 'sel\_txt' items). Below this main flow, a 'Dynamic procedure' section shows a 'Fine-tuning' graph with a 'loss' metric and a 'Content generation' sequence of 'gen\_txt' items.

54

**Perceptual Analysis**

**Frontiers: Determining the Validity of Large Language Models for Automated Perceptual Analysis**

Pelyao Li,<sup>a</sup> Noah Castelo,<sup>b</sup> Zsolt Katona,<sup>a,\*</sup> Miklos Sarvary<sup>c</sup>

<sup>a</sup> Haas School of Business, University of California, Berkeley, California 94720; <sup>b</sup> Alberta School of Business, University of Alberta, Edmonton, Alberta T6G 2R6, Canada; <sup>c</sup> Columbia Business School, Columbia University, New York, New York 10027

\*Corresponding author. Contact: ophiklsj@berkeley.edu (PL); ncastelo@ualberta.ca (NC); zskatona@haas.berkeley.edu; https://orcid.org/0000-0001-8411-6040 (ZK); miklos.sarvary@columbia.edu; https://orcid.org/0000-0002-3801-5917 (MS)

Received: September 28, 2023  
Revised: November 7, 2023  
Accepted: December 5, 2023  
Published Online in Articles in Advance: January 25, 2024

<https://doi.org/10.1287/mksc.2023.0454>  
Copyright © 2024 INFORMS

**Abstract:** This paper explores the potential of large language models (LLMs) to substitute for human participants in market research. Such LLMs can be used to generate test text given a prompt. We argue that perceptual analysis is a naturally proposed use case for such automated market research for certain product categories. We propose a methodology that generates outputs that closely match those generated from human surveys: agreement rates between human- and LLM-generated data sets reach over 75%. Moreover, this applies for perceptual analysis based on both brand similarity measures and product attribute ratings. The paper demonstrates that, for some categories, this new method of fully or partially automated market research will increase the efficiency of market research by meaningfully speeding up the process and potentially reducing the cost. Further results also suggest that with an ever larger training corpus applied to large language models, LLM-based market research will be applicable to answer more nuanced questions based on demographic variables or contextual variation that would be prohibitively expensive or infeasible with human respondents.

**History:** Catherine Tucker served as the senior editor. This paper was accepted through the *Marketing Science* Frontiers review process.  
**Funding:** This work was supported by the Social Sciences and Humanities Research Council of Canada [Grant A30-2021-0003].  
**Supplemental Material:** The online appendix and data files are available at <https://doi.org/10.1287/mksc.2023.0454>.

**Keywords:** artificial intelligence • perceptual maps • large language model • natural language processing • market research

**Frontiers: Determining the Validity of Large Language Models for Automated Perceptual Analysis**

P Li, N Castelo, Z Katona, M Sarvary - Marketing Science, 2024 - pubsonline.informs.org

... This paper explores the potential of **large language models** (LLMs) to substitute for human ... an ever larger training corpus applied to **large language models**, LLM-based market research ...

☆ Save 95 Cite Cited by 2 Related articles ☰

55

**LLMs as Simulators for Humans**

**Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?**

John J. Horton

WORKING PAPER 31122 DOI 10.3386/w31122 ISSUE DATE April 2023

Newly-developed large language models (LLMs)—because of how they are trained and designed—are implicit computational models of humans—a homo silicus. LLMs can be used like economists use homo economicus: they can be given endowments, information, preferences, and so on, and then their behavior can be explored in scenarios via simulation. Experiments using this approach, derived from Charness and Rabin (2002), Kahneman, Knetsch and Thaler (1986), and Samuelson and Zeckhauser (1988) show qualitatively similar results to the original, but it is also easy to try variations for fresh insights. LLMs could allow researchers to pilot studies via simulation first, searching for novel social science insights to test in the real world.

Using large language models to simulate multiple humans and replicate human subject studies

GV Aher, RI Arriaga, AT Kalai

International Conference on Machine Learning, 2023 - proceedings.mlr.press

**Abstract**

We introduce a new type of test, called a Turing Experiment (TE), for evaluating to what extent a given language model, such as GPT models, can simulate different aspects of human behavior. A TE can also reveal consistent distortions in a language model's simulation of a specific human behavior. Unlike the Turing Test, which involves simulating a single arbitrary individual, a TE requires simulating a representative sample of participants in human subject research. We carry out TEs that attempt to replicate well-known results from human subjects.

SHOW MORE ☰

☆ Save 95 Cite Cited by 159 Related articles All 8 versions ☰

**Large language models as simulated economic agents: What can we learn from homo silicus?**

JJ Horton  
2023 - nber.org

**Abstract**

Newly-developed large language models (LLMs)—because of how they are trained and designed—are implicit computational models of humans—a homo silicus. LLMs can be used like economists use homo economicus: they can be given endowments, information, preferences, and so on, and then their behavior can be explored in scenarios via simulation. Experiments using this approach, derived from Charness and Rabin (2002), Kahneman, Knetsch and Thaler (1986), and Samuelson and Zeckhauser (1988) show qualitatively similar results to the original, but it is also easy to try variations for fresh insights. LLMs could allow researchers to pilot studies via simulation first, searching for novel social science insights to test in the real world.

SHOW MORE ☰

☆ Save 95 Cite Cited by 150 Related articles All 11 versions ☰

56

# LLMs as Simulators for Causal Inference

Automated Social Science:  
A Structural Causal Model-Based Approach\*

Benjamin S. Manning<sup>†</sup> Kehang Zhu<sup>†</sup> John J. Horton  
MIT Harvard MIT & NBER

March 13, 2024

The Challenge of Using LLMs to Simulate Human Behavior:  
A Causal Inference Perspective

George Gui and Olivier Toubia\*

December 24, 2023

## Abstract

We present an approach for automatically generating and testing, *in silico*, social scientific hypotheses. This automation is made possible by recent advances in large language models (LLM), but the key feature of the approach is the use of structural causal models. Structural causal models provide a language to state hypotheses, a blueprint for constructing LLM-based agents, an experimental design, and a plan for data analysis. The fitted structural causal model becomes an object available for prediction or the planning of follow-on experiments. We demonstrate the approach with several scenarios: a negotiation, a bail hearing, a job interview, and an auction. In each case, causal relationships are proposed and tested, finding evidence for some and not others. In the auction experiment, we show that the *in silico* simulation results closely match the predictions of auction theory, but elicited predictions of the clearing prices from an LLM are inaccurate. However, the LLM's predictions are dramatically improved if the model can condition on the fitted structural causal model. When given a proposed structural causal model for each of the scenarios, the LLM is good at predicting the signs of estimated effects, but it cannot reliably predict the magnitudes of those estimates. This suggests that simulations of social interactions give the model insight not available purely through direct elicitation. In short, the LLM knows more than it can (immediately) tell.

## Abstract

Large Language Models (LLMs) have demonstrated impressive potential to simulate human behavior. Using a causal inference framework, we empirically and theoretically analyze the challenges of conducting LLM-simulated experiments, and explore potential solutions. In the context of demand estimation, we show that variations in the treatment included in the prompt (e.g., price of focal product) can cause variations in unspecified confounding factors (e.g., price of competitors, historical prices, outside temperature), introducing endogeneity and yielding implausibly flat demand curves. We propose a theoretical framework suggesting this endogeneity issue generalizes to other contexts and won't be fully resolved by merely improving the training data. Unlike real experiments where researchers assign pre-existing units across conditions, LLMs simulate units based on the entire prompt, which includes the description of the treatment. Therefore, due to associations in the training data, the characteristics of individuals and environments simulated by the LLM can be affected by the treatment assignment. We explore two potential solutions. The first specifies all contextual variables that affect both treatment and outcome, which we demonstrate to be challenging for a general-purpose LLM. The second explicitly specifies the source of treatment variation in the prompt given to the LLM (e.g., by informing the LLM that the store is running an experiment). While this approach only allows the estimation of a conditional average treatment effect that depends on the specific experimental design, it provides valuable directional results for exploratory analysis.

57

57

# LLMs Approximate Humans

## A Turing test of whether AI chatbots are behaviorally similar to humans

[Q Mei](#), [Y Xie](#), [W Yuan](#), [MO Jackson](#)

Proceedings of the National Academy of Sciences, 2024 • National Acad Sciences

We administer a Turing test to AI chatbots. We examine how chatbots behave in a suite of classic behavioral games that are designed to elicit characteristics such as trust, fairness, risk-aversion, cooperation, etc., as well as how they respond to a traditional Big-5 psychological survey that measures personality traits. ChatGPT-4 exhibits behavioral and personality traits that are statistically indistinguishable from a random human from tens of thousands of human subjects from more than 50 countries. Chatbots also modify their behavior based on previous experience and contexts “as if” they were learning from the interactions and change their behavior in response to different framings of the same strategic situation. Their behaviors are often distinct from average and modal human behaviors, in which case they tend to behave on the more altruistic and cooperative end of the distribution. We estimate that they act as if they are maximizing an average of their own and partner’s payoffs.

National Acad Sciences

收起 ^

☆ 保存 引用 被引用次数: 7 相关文章 所有 4 个版本

- GPT-4 exhibit behavioral and personality traits that are statistically indistinguishable from a random human.
- LLMs “learn” from previous interactions, in response to experiences and contexts.
- LLMs are tested using classic behavioral games in psychology.

58

58

## Artificial Agents and Operations Management Decision-Making

S Kirshner

Available at SSRN 4726933, 2024 · papers.ssrn.com

### Abstract

Large language models (LLMs) like GPT are increasingly being integrated into organizational workflows to improve operational processes and enhance efficiencies. To provide insights into these advancements, this study examines GPT within the context of operations management (OM) to understand how the LLM solves OM problems and how its decision processes compare with humans. To evaluate GPT's OM decision-making capabilities, this research conducts 9 experiments from the recent MS Replication Project, utilizing state-of-the-art custom GPT assistants as participants. The performance of the assistants is assessed against the key hypotheses identified in the replication projects to determine whether GPT assistants behave optimally or align with human behavior. Interestingly, in 7 of 9 experiments, assistants' responses were consistent with human decision-making. Collectively examining the results across the suite of studies reveals rich insights into the decision strategies of GPT agents. For instance, in studies involving other decision-makers, the assistants employed satisfying heuristics instead of comprehensively considering others' strategies and optimizing accordingly. In experiments without other decision-makers, GPT often relied on moving average forecasts, which led to pulled-to-center orders and system neglect. This research extends the emerging GPT experimental literature by considering more complex decision-making scenarios and unearthing novel decision processes. The findings also provide many exciting directions for future research by underscoring the need for additional GPT OM-based experiments and the importance of developing interventions for LLMs solving OM problems.

[papers.ssrn.com](#)

收起 ^

☆ 保存 ⚡ 引用 相关文章 ⟲

Kremer and Debo (2016) Queuing

The presence of some informed consumers makes uninformed consumers a) less likely to purchase given a short wait time and b) less sensitive to wait time increases.

Croson and Donohue (2006)

☒ GPT does not change its decision-making strategy given the presence of informed consumers despite recognizing their potential informational value.

Supply Chains

Sharing dynamic inventory information across the supply chain will decrease the level of order oscillation.

☒ GPT's order oscillations dramatically increased with dynamic inventory information, contrary to the hypothesis.

Two unsuccessful replications by GPT.

59

## LLMs Replicate Behavioral OM Studies

- *Management Science* conducted a replication study of 9 classic results in BOM. Can LLMs replicate these studies?
- LLMs could replicate 7 out of 9 BOM classic papers.
- Satisfying heuristics and pulling-to-the-center behaviors exhibit.

59

## [PDF] Synthetic replacements for human survey data? the perils of large language models

J Bisbee, J Clinton, C Dorff, B Kenkel, J Larson

SocArXiv. May, 2023 · scholar.archive.org

### Abstract

Large Language Models (LLMs) offer new research possibilities for social scientists, but their potential as "synthetic data" is still largely unknown. In this paper, we investigate how accurately the popular closed-source LLM ChatGPT can recover public opinion, prompting the LLM to adopt different "personas" and then provide feeling thermometer scores for 11 sociopolitical groups. The average scores generated by ChatGPT correspond closely to the averages in our baseline survey, the 2016–2020 American National Election Study. Nevertheless, sampling by ChatGPT is not reliable for statistical inference: there is less variation in responses than in the real surveys, and regression coefficients often differ significantly from equivalent estimates obtained using ANES data. We also document how the distribution of synthetic responses varies with minor changes in prompt wording, and we show how the same model yields significantly different results over a three-month period. Altogether, our findings raise serious concerns about the quality, reliability, and reproducibility of synthetic survey data generated by LLMs.

[scholar.archive.org](#)

收起 ^

☆ 保存 ⚡ 引用 被引用次数: 6 相关文章 ⟲

Forthcoming at *Political Analysis*

## When are LLMs not Reliable in Political Surveys?

- LLMs are prompted to answer political survey questions without any contexts.
- LLMs behave more extremely and have less variations than real human surveys.
- Data collected from LLMs may not be reliable for statistical inference.

60

60

## LLM-Human Collaborations

Large language model in creative work: The role of collaboration modality and user expertise

Z Chen, J Chan

Available at SSRN 4575598, 2023 · papers.ssrn.com

### Abstract

Since the launch of ChatGPT in Dec 2022, Large Language Models (LLMs) are rapidly adopted by businesses to assist users in a wide range of open-ended tasks, including those that require creativity. While the versatility of LLM has unlocked new ways of human-AI collaboration, it remains uncertain whether LLMs can truly enhance business outcomes. To examine the effects of human-LLM collaboration on business outcomes, we conducted an experiment where we tasked expert and non-expert users to write an ad copy with and without the assistance of LLMs. Here, we investigate and compare two ways of working with LLMs:(1) using LLMs as "ghostwriters," which assume the main role of content generation task and (2) using LLMs as "sounding boards," to provide feedback on human-created content. We measure the quality of the ads using the number of clicks generated by the created ads on major social media platforms. Our results show that different collaboration modalities can result in very different outcomes for different user types. Using LLMs as sounding boards enhances the quality of the resultant ad copies, especially for non-experts. However, using LLMs as ghostwriters did not provide significant benefits and is in fact detrimental to expert users. We rely on textual analyses to understand the mechanisms and learned that using LLMs as ghostwriters produces an anchoring effect which leads to lower-quality ads. On the other hand, using LLMs as sounding boards helped non-experts achieve ad content with low semantic divergence to content produced by experts, thereby closing the gap between the two types of users.

[papers.ssrn.com](#)

SHOW LESS ^

☆ Save 99 Cite Cited by 8 Related articles ▾

61

- LLM replaces your work or refines your work?
- LLM refines your work: Quality improves especially for non-experts.
- LLM replaces your work: Quality does not improve and decreases for experts.
- Mechanism:
  - Replacement: Anchoring effect
  - Refinement: Low semantic divergence between experts and nonexperts.

61

## LLMs in Other Contexts

[HTML] Highly accurate protein structure prediction with AlphaFold

J Jumper, R Evans, A Pritzel, T Green, M Figurnov, O Ronneberger, K Tunyasuvunakool...

Nature, 2021 · nature.com

### Abstract

Proteins are essential to life, and understanding their structure can facilitate a mechanistic understanding of their function. Through an enormous experimental effort 1, 2, 3, 4, the structures of around 100,000 unique proteins have been determined 5, but this represents a small fraction of the billions of known protein sequences 6, 7. Structural coverage is bottlenecked by the months to years of painstaking effort required to determine a single protein structure. Accurate computational approaches are needed to address this gap and to enable large-scale structural bioinformatics. Predicting the three-dimensional structure that a protein will adopt based solely on its amino acid sequence—the structure prediction component of the ‘protein folding problem’—has been an important open research problem for more than 50 years 9. Despite recent progress 10, 11, 12, 13, 14, existing methods fall far short of atomic accuracy, especially when no homologous structure is available. Here we provide the first computational method that can regularly predict protein structures with atomic accuracy even in cases in which no similar structure is known. We validated an entirely redesigned version of our neural network-based model, AlphaFold, in the challenging 14th Critical Assessment of protein Structure Prediction (CASP14) 15, demonstrating accuracy competitive with experimental structures in a majority of cases and greatly outperforming other methods. Underpinning the latest version of AlphaFold is a novel machine learning approach that incorporates physical and biological knowledge about protein structure, leveraging multi-sequence alignments, into the design of the deep learning algorithm.

[nature.com](#)

SHOW LESS ^

☆ Save 99 Cite Cited by 20650 Related articles All 25 versions Web of Science: 11921 ▾

62

A decoder-only foundation model for time-series forecasting

A Das, W Kong, R Sen, Y Zhou

arXiv preprint arXiv:2310.10688, 2023 · arxiv.org

Motivated by recent advances in large language models for Natural Language Processing (NLP), we design a time-series foundation model for forecasting whose out-of-the-box zero-shot performance on a variety of public datasets comes close to the accuracy of state-of-the-art supervised forecasting models for each individual dataset. Our model is based on pretraining a patched-decoder style attention model on a large time-series corpus, and can work well across different forecasting history lengths, prediction lengths and temporal granularities.

[arxiv.org](#)

SHOW LESS ^

☆ Save 99 Cite Cited by 16 Related articles All 2 versions ▾

62