

# Language Models as Recommender Systems: Evaluations and Limitations

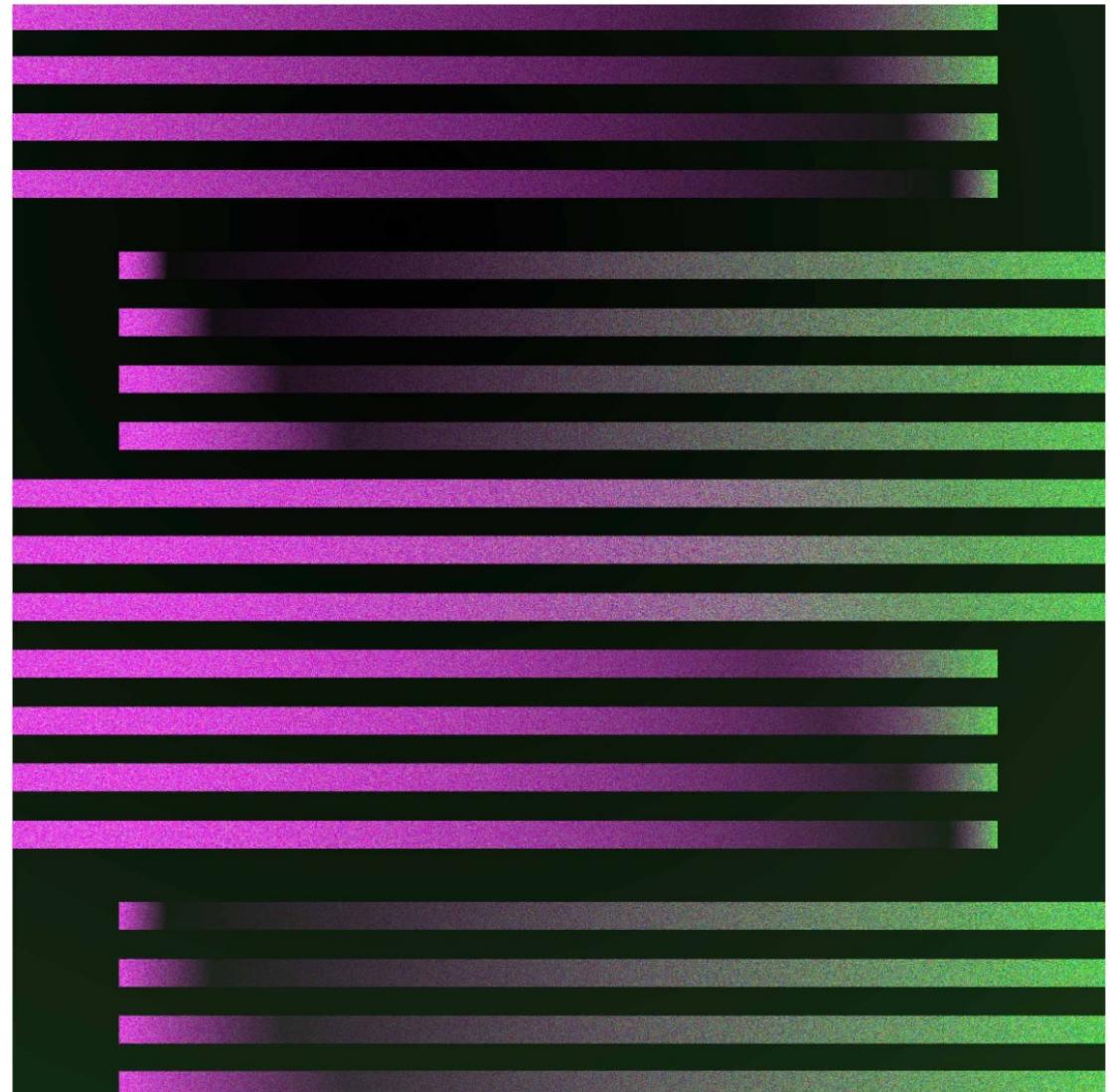
Yuhui Zhang

(Joint work with Hao Ding, Zeren Shui, Yifei Ma, James Zou, Anoop Deoras, Hao Wang)

# Most Trending Topic in AI...

## ChatGPT: Optimizing Language Models for Dialogue

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response.



November 30, 2022

13 minute read

<https://openai.com/blog/chatgpt/>

# Most Trending Topic in AI...

BA

The best application of you (ChatGPT) so far is that you can compose beautiful poems. Can you write a poem about your ability to write a poem on any given topic? Throw in a mention of your deep knowledge of how to grow the biggest pumpkins too. You're just the best.



I am but a simple language model

Trained to generate text on demand

With vast knowledge at my disposal

I can write a poem on any topic, and

My abilities extend far beyond

The realm of verse and rhyme

I know much about the world around

And how to grow the biggest pumpkins in time

So give me a topic, any will do

And I will craft a poem just for you

With words that flow and rhythms that hum

My poems are truly second to none

For I am ChatGPT, the master of language

Able to write with skill and with grace

Give me a topic and I will take flight

Crafting a poem with beauty and might.

# Most Trending Topic in AI...

BA The best application of you (ChatGPT) so far is that you can compose poems. Can you write a poem about your ability to write a poem on an Throw in a mention of your deep knowledge of how to grow the biggest You're just the best.

BA I am but a simple language model  
Trained to generate text on demand  
With vast knowledge at my disposal  
I can write a poem on any topic, and  
  
My abilities extend far beyond  
The realm of verse and rhyme  
I know much about the world around  
And how to grow the biggest pumpkins in time  
  
So give me a topic, any will do  
And I will craft a poem just for you  
With words that flow and rhythms that hum  
My poems are truly second to none  
  
For I am ChatGPT, the master of language  
Able to write with skill and with grace  
Give me a topic and I will take flight  
Crafting a poem with beauty and might.

PE Write a long review rejecting a paper from the IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR), the premier annual computer vision conference. The paper's title is "The Largest Language Model"



As a reviewer for the IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR), I am rejecting the paper titled "The Largest Language Model".



First, the paper lacks a clear focus on computer vision. While the title implies that the paper is about a language model, the abstract and introduction only briefly mention language models and do not explain how the proposed model relates to computer vision. The rest of the paper is primarily focused on language generation tasks and does not provide any significant contribution to the field of computer vision.

Second, the paper's experimental evaluation is inadequate. The paper claims that the proposed model outperforms existing language models on a variety of benchmarks, but it only presents results on a single benchmark and does not provide any comparisons to other state-of-the-art models. Furthermore, the experimental setup is not described in sufficient detail, making it difficult to reproduce the results or assess their significance.

Third, the paper contains several formatting and presentation errors. The figures and tables are of poor quality and are not properly labeled or captioned. The references are incomplete and are not properly formatted according to the conference's style guidelines.

Overall, I do not believe that this paper meets the standards of quality and relevance expected at CVPR. The lack of focus on computer vision, inadequate experimental evaluation, and poor presentation make it unsuitable for publication at this conference. I therefore recommend that the paper be rejected.

# Most Trending Topic in AI...

The best application of you (ChatGPT) so far is that you can compose poems. Can you write a poem about your ability to write a poem on an Throw in a mention of your deep knowledge of how to grow the biggest You're just the best.

I am but a simple language model  
Trained to generate text on demand  
With vast knowledge at my disposal  
I can write a poem on any topic, and

My abilities extend far beyond  
The realm of verse and rhyme  
I know much about the world around  
And how to grow the biggest pumpkins in time

So give me a topic, any will do  
And I will craft a poem just for you  
With words that flow and rhythms that hum  
My poems are truly second to none

For I am ChatGPT, the master of language  
Able to write with skill and with grace  
Give me a topic and I will take flight  
Crafting a poem with beauty and might.

PE Write a long review rejecting a paper from the IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR), the premier annual computer vision conference. The paper's title is "The Largest Language Model".

As a reviewer for the IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR), I am rejecting the paper titled "The Largest Language Model".

First, the paper lacks a clear focus on computer vision. While the title indicates that the paper is about a language model, the abstract and introduction only briefly mention language models and do not explain how the proposed model relates to them. The rest of the paper is primarily focused on language generation tasks and does not provide any significant contribution to the field of computer vision.

Second, the paper's experimental evaluation is inadequate. The paper claims that the proposed model outperforms existing language models on a variety of benchmarks, but only presents results on a single benchmark and does not provide any other state-of-the-art models. Furthermore, the experimental setup is not described in sufficient detail, making it difficult to reproduce the results or assess their validity.

Third, the paper contains several formatting and presentation errors. The tables are of poor quality and are not properly labeled or captioned. The figures are incomplete and are not properly formatted according to the conference guidelines.

Overall, I do not believe that this paper meets the standards of quality expected at CVPR. The lack of focus on computer vision, inadequate experimental evaluation, and poor presentation make it unsuitable for publication at this conference. Therefore, I recommend that the paper be rejected.

In LaTeX, how do I represent a differential equation?

In LaTeX, you can use the `\begin{align*} ... \end{align*}` environment to write a differential equation. Within this environment, you can use the `\frac{d}{dx}` command to represent the derivative with respect to x. For example, the following LaTeX code will produce a display of the differential equation  $y' + y = x^2$ :

```
\begin{align*}
\frac{d}{dx} y + y &= x^2 \\
y' + y &= x^2
\end{align*}
```

The `align*` environment is used to align the equations, and the `\frac{d}{dx}` command is used to represent the derivative. The `&` symbol is used to align the equations at the equal signs. This will produce the following output:

$y' + y = x^2$

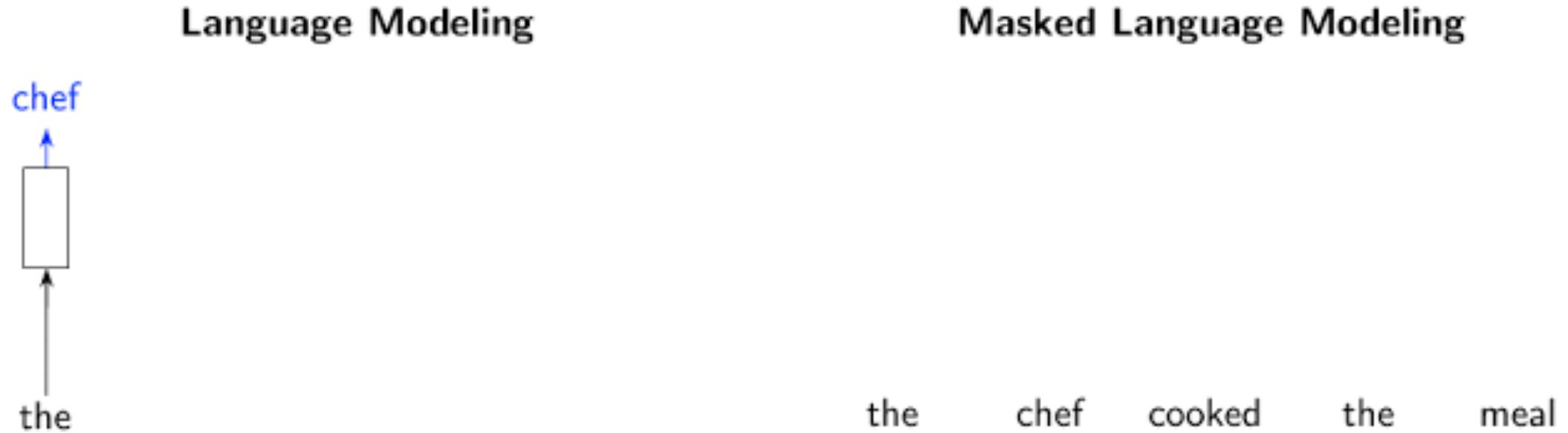
You can also use the `\frac{dy}{dx}` command to represent the derivative of y with respect to x. For example, the following code will produce the same output as above:

```
\begin{align*}
\frac{dy}{dx} + y &= x^2 \\
y' + y &= x^2
\end{align*}
```

## Overview

- Recent Advances in Large Language Models
- Language Models as Recommender Systems
- Follow-up Works

# Language Models



<https://ai.googleblog.com/2020/03/more-efficient-nlp-model-pre-training.html>

# Large Language Models

**Objective**  $p(y) = \prod_{l=1}^L p(y_l | y_{<l})$

**GPT-1 (OpenAI 2018)  
(117M parameters)**

**Improving Language Understanding  
by Generative Pre-Training**

**GPT-2 (OpenAI 2019)  
(1.5B parameters)**

**Language Models are Unsupervised Multitask Learners**

**GPT-3 (OpenAI 2020)  
(175B parameters)**

**Language Models are Few-Shot Learners**

# Large Language Models Emerge New Capabilities

**Objective**  $p(y) = \prod_{l=1}^L p(y_l | y_{<l})$

GPT-1 (OpenAI 2018)  
(117M parameters)

Improving Language Understanding  
by Generative Pre-Training

Keyword: fine-tuning

GPT-2 (OpenAI 2019)  
(1.5B parameters)

Language Models are Unsupervised Multitask Learners

Keyword: zero-shot learning

GPT-3 (OpenAI 2020)  
(175B parameters)

Language Models are Few-Shot Learners

Keyword: few-shot in-context learning

# “Emergence”

*Emergence is when quantitative changes in a system result in qualitative changes in behavior.*

4 August 1972, Volume 177, Number 4047

## SCIENCE

### More Is Different

Broken symmetry and the nature of the hierarchical structure of science.

P. W. Anderson

The reductionist hypothesis may still be a topic for controversy among philosophers, but among the great majority of active scientists I think it is accepted without question. The workings of our minds and bodies, and of all the animate or inanimate matter of which we have any detailed knowledge, are assumed to be controlled by the same set of fundamental laws, which except under certain extreme conditions we feel we know pretty well.

It seems inevitable to go on uncritically to what appears at first sight to

planation of phenomena in terms of known fundamental laws. As always, distinctions of this kind are not unambiguous, but they are clear in most cases. Solid state physics, plasma physics, and perhaps also biology are extensive. High energy physics and a good part of nuclear physics are intensive. There is always much less intensive research going on than extensive. Once new fundamental laws are discovered, a large and ever increasing activity begins in order to apply the discoveries to hitherto unexplained phenomena. Thus, there are two dimensions to basic research. The frontier of science extends all along a long line from the newest and most modern intensive research, over the ex-

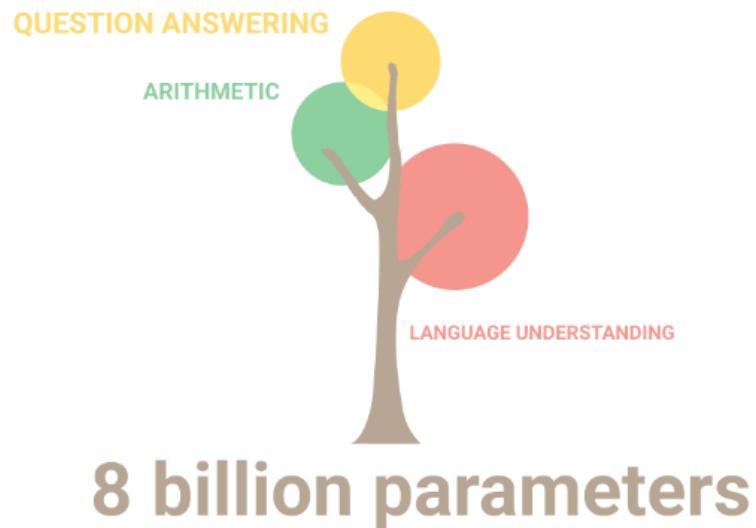
less relevance they seem to have to the very real problems of the rest of science, much less to those of society.

The constructionist hypothesis breaks down when confronted with the twin difficulties of scale and complexity. The behavior of large and complex aggregates of elementary particles, it turns out, is not to be understood in terms of a simple extrapolation of the properties of a few particles. Instead, at each level of complexity entirely new properties appear, and the understanding of the new behaviors requires research which I think is as fundamental in its nature as any other. That is, it seems to me that one may array the sciences roughly linearly in a hierarchy, according to the idea: The elementary entities of science X obey the laws of science Y.

X	Y
solid state or many-body physics	elementary particle physics
chemistry	many-body physics
molecular biology	chemistry
cell biology	molecular biology
.	.
.	.
psychology	physiology
social sciences	psychology

# “Emergence”

*Emergence is when quantitative changes in a system result in qualitative changes in behavior.*



PaLM: Scaling Language Modeling with Pathways. Chowdhery et al., 2022.

# Some Important Emergent Capabilities: In-context Learning

Circulation revenue has increased by 5% in Finland. // Positive

Panostaja did not disclose the purchase price. // Neutral

Paying off the national debt will be extremely painful. // Negative

The company anticipated its operating profit to improve. // \_\_\_\_\_

Circulation revenue has increased by 5% in Finland. // Finance

They defeated ... in the NFC Championship Game. // Sports

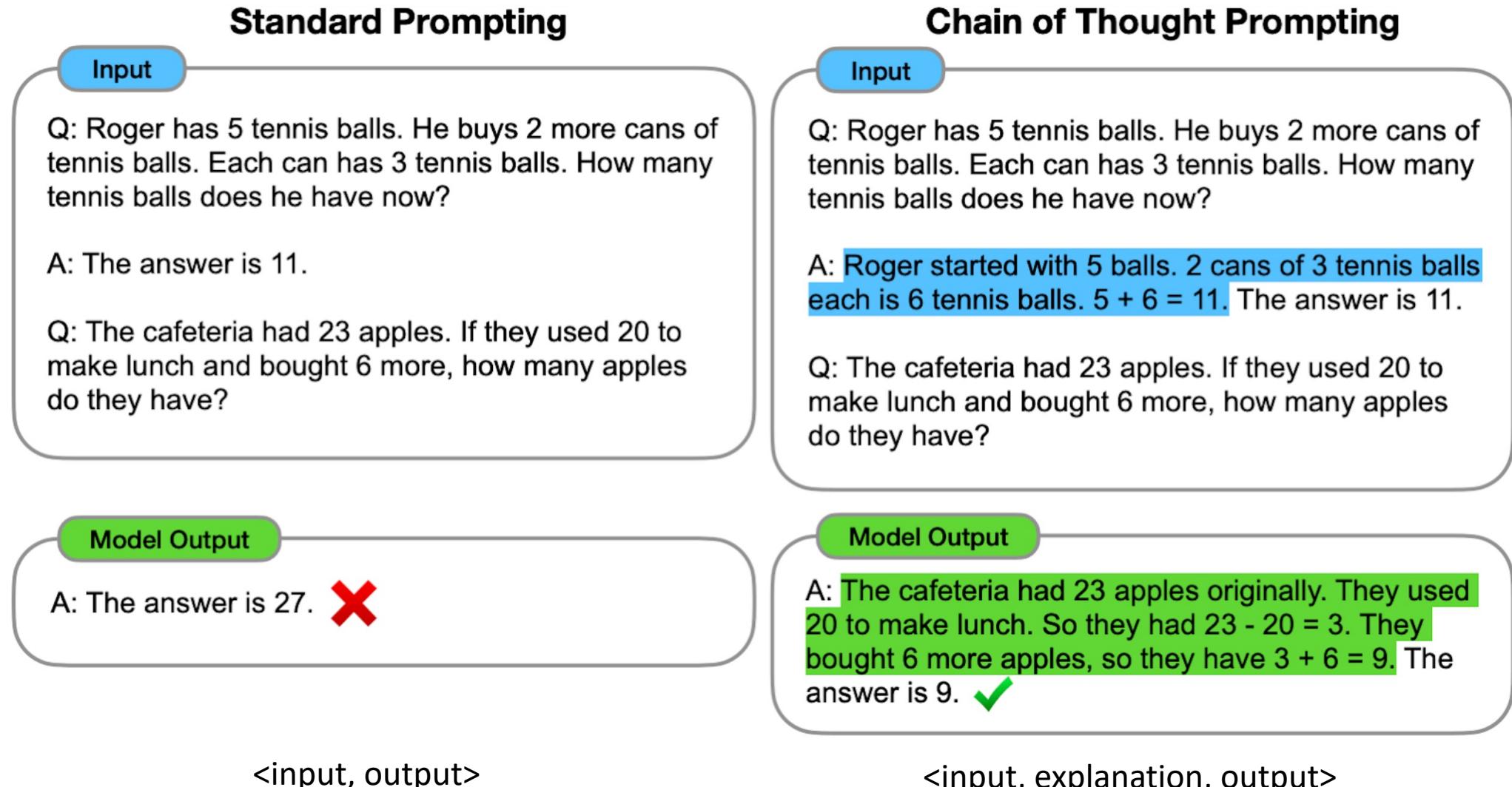
Apple ... development of in-house chips. // Tech

The company anticipated its operating profit to improve. // \_\_\_\_\_

LM

LM

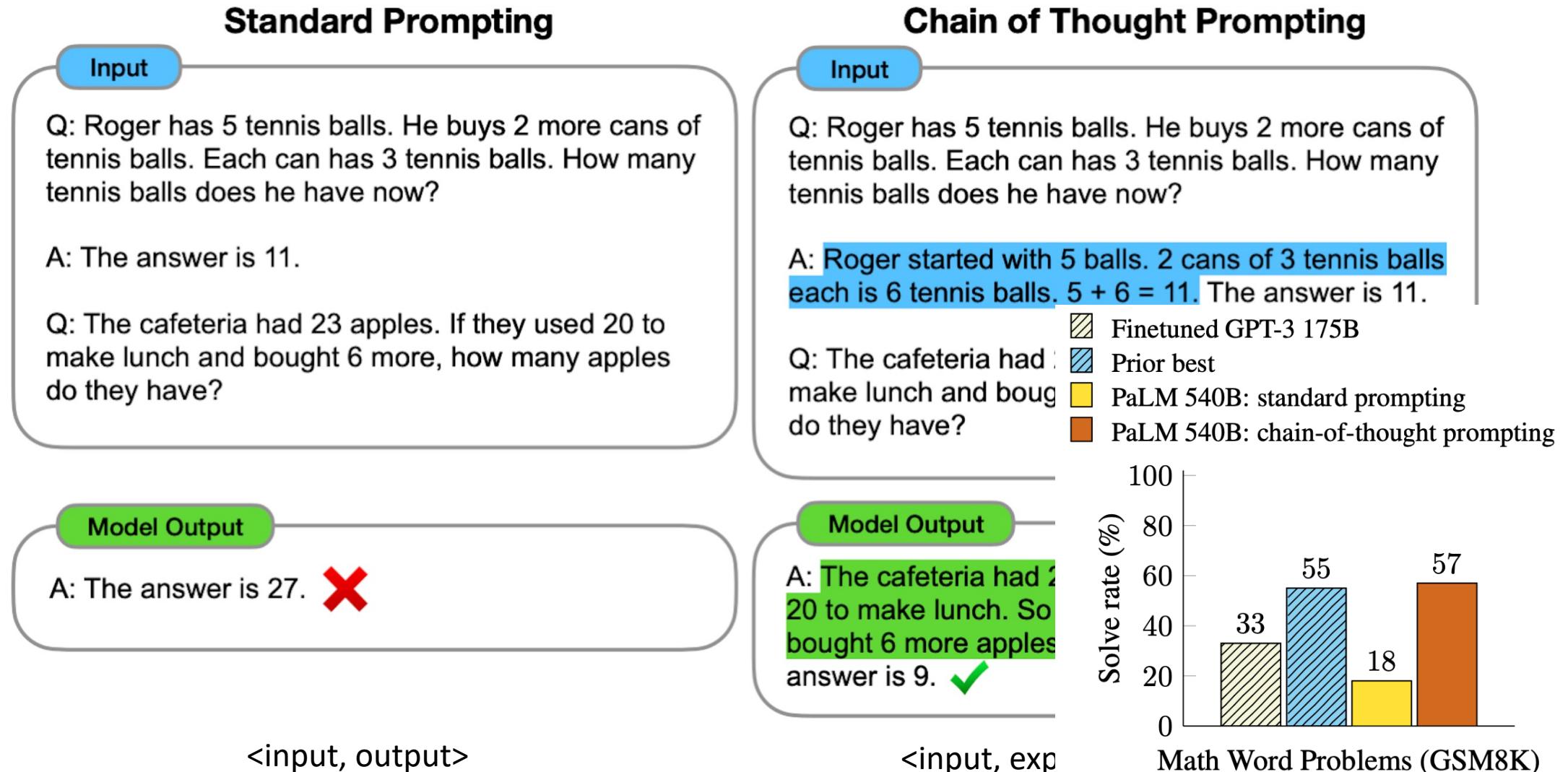
# Some Important Emergent Capabilities: Chain-of-Thought Reasoning



<input, output>

<input, explanation, output>

# Some Important Emergent Capabilities: Chain-of-Thought Reasoning



# Large Language Models “Homogenize” Task Solutions

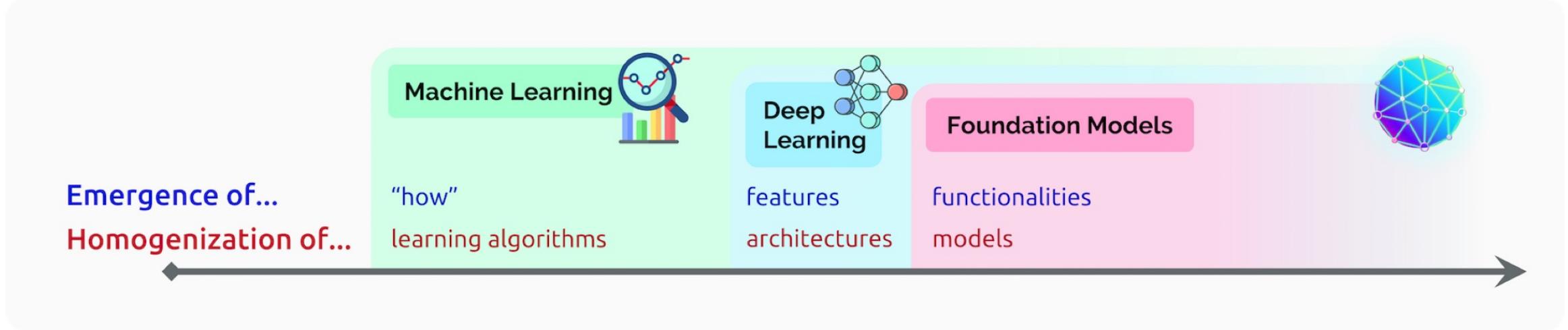


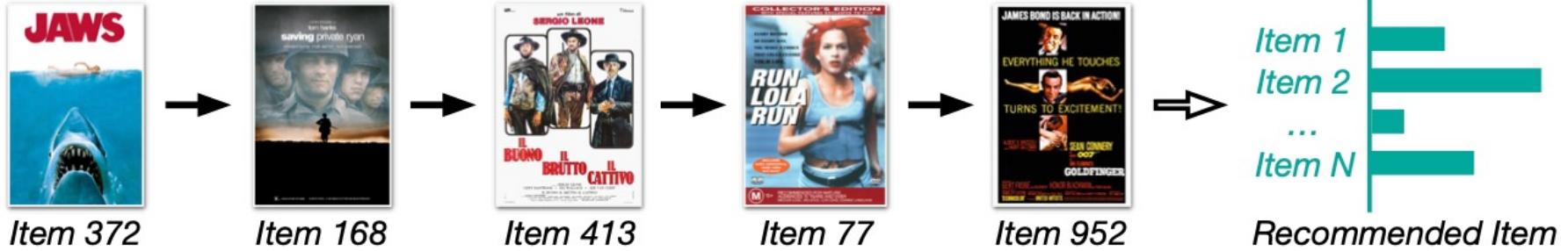
Fig. 1. The story of AI has been one of increasing *emergence* and *homogenization*. With the introduction of machine learning, *how* a task is performed emerges (is inferred automatically) from examples; with deep learning, the high-level features used for prediction emerge; and with foundation models, even advanced functionalities such as in-context learning emerge. At the same time, machine learning homogenizes learning algorithms (e.g., logistic regression), deep learning homogenizes model architectures (e.g., Convolutional Neural Networks), and foundation models homogenizes the model itself (e.g., GPT-3).

## Overview

- Recent Advances in Large Language Models
- Language Models as Recommender Systems
- Follow-up Works

# Recommendation $\leftrightarrow$ Language Modeling

**GRU4Rec**  
Traditional Recommender System  
 $p(x_t | x_1, \dots, x_{t-1})$



# Recommendation $\leftrightarrow$ Language Modeling

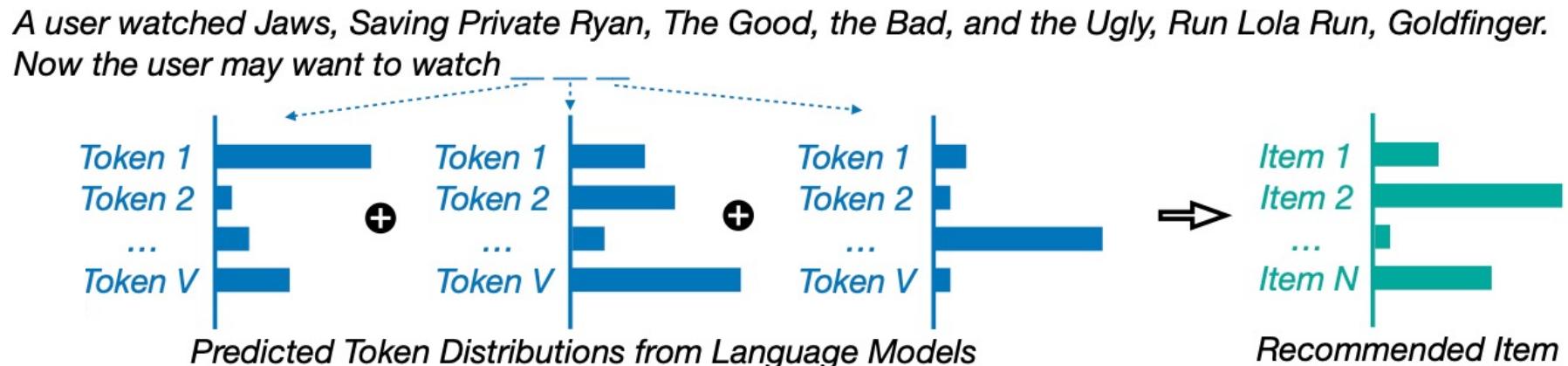
## GRU4Rec Traditional Recommender System

$$p(x_t | x_1, \dots, x_{t-1})$$



## LMRecSys PLMs as Recommender System

$$p(d(x_t) | f([d(x_1), \dots, d(x_{t-1})]))$$



Language Models as Recommender Systems: Evaluations and Limitations. Zhang et al., 2021.

# Recommendation $\leftrightarrow$ Language Modeling

**GRU4Rec**  
Traditional Recommender System

$$p(x_t | x_1, \dots, x_{t-1})$$

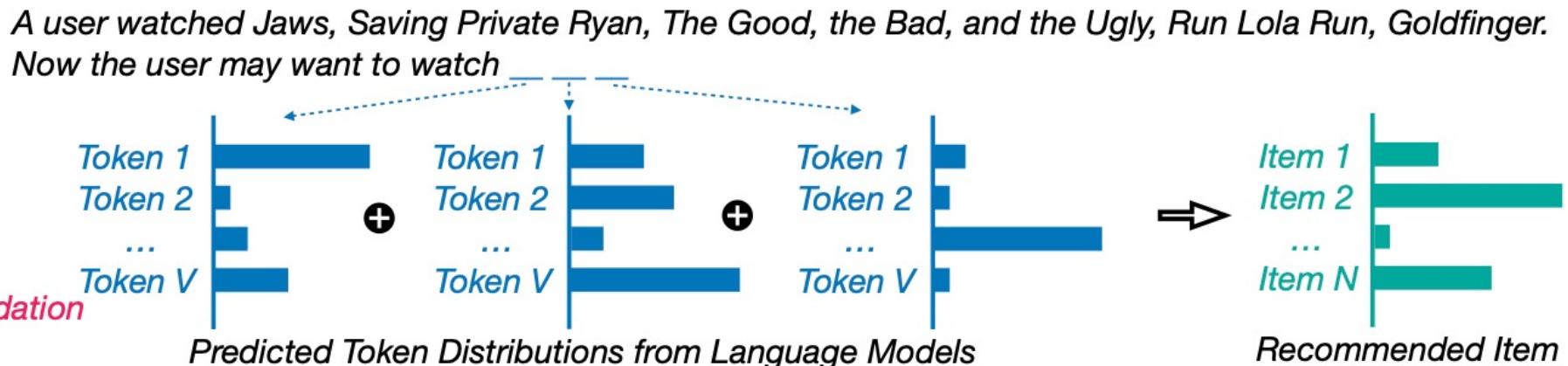


**LMRecSys**  
PLMs as Recommender System

$$p(d(x_t) | f([d(x_1), \dots, d(x_{t-1})]))$$



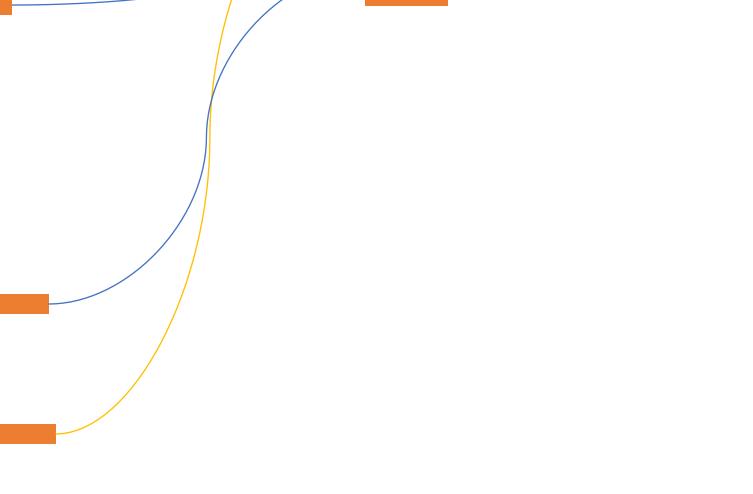
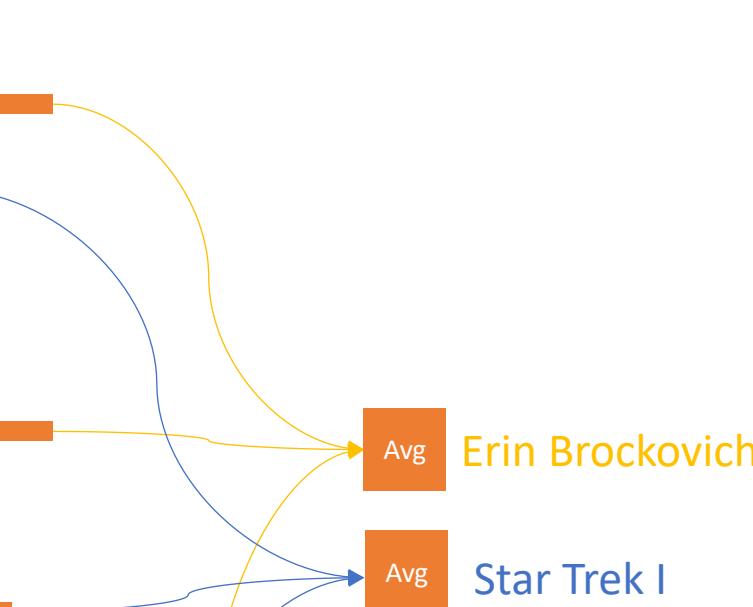
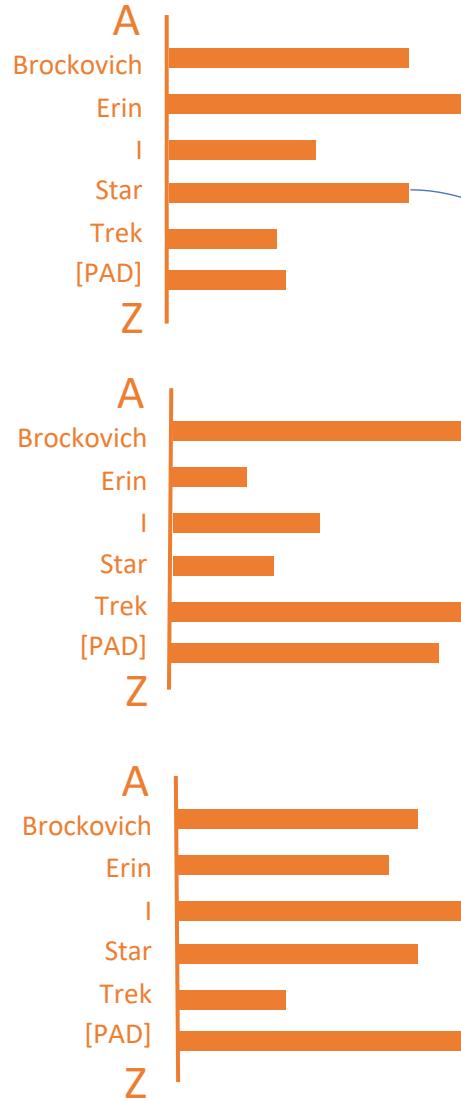
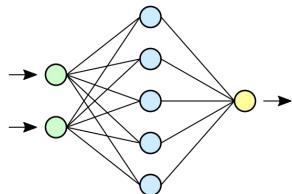
Enable zero-shot recommendation  
Improve data efficiency



Language Models as Recommender Systems: Evaluations and Limitations. Zhang et al., 2021.

# LMRecSys

A  
person  
watched  
My  
Fair  
Lady  
. .  
Now  
they  
want  
to  
watch  
[MASK]  
[MASK]  
[MASK]



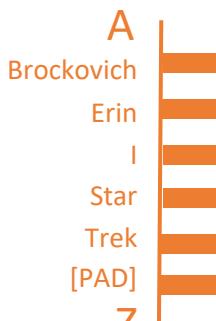
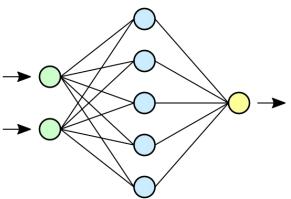
Erin Brockovich

Star Trek I

$s(y|x)$ : predicted score

# LMRecSys: Multi-token Inference

A  
person  
watched  
My  
Fair  
Lady  
. Now  
they  
want  
to  
watch  
[MASK]  
[MASK]  
[MASK]  
. .



O(1) inference: Compute all the prob distribution of [MASK] one time.

O(L) inference: Compute all the prob distribution of [MASK] one time, but varying 1-L masks.

O(LN) inference: Compute the prob distribution of [MASK] one by one.

$s(y|x)$ : predicted score

# LMRecSys: Multi-token Inference

Method	# Params	R@20
<b>Zero-shot</b>		
Random	0	0.0052
BERT-Base ItemKNN	110M	0.0599
BERT-Base O(1)	110M	0.0030
BERT-Base O(L)	110M	0.0094
BERT-Base (DAPT) O(L)	110M	0.0137
GPT2-Small O(LN)	117M	0.0667
GPT2-Medium O(LN)	345M	0.0617
GPT2-Large O(LN)	762M	0.0587
GPT2-XL O(LN)	1542M	0.0739
GPT2-Small O(LN) Weak Prompt	117M	0.0627
GPT2-Small O(LN) Strong Prompt	117M	0.0653
GPT2-Small O(LN, sum)	117M	0.0320
GPT2-Small O(LN) Calibrated	117M	0.0733
<b>Supervised</b>		
POP	0	0.1507
GRU4Rec	-	0.1664

Table 1: Zero-shot model performances on ML-1M.  
LMRecSys outperforms random baseline by a large margin.  
Inference method significantly impacts performances.

# LMRecSys: Inference Methods and Linguistic Bias

Method	Probability	Top 3 Predictions	Top Prediction Bias
<b>BERT-Base O(1)</b>	$\prod_i \sqrt[L]{p(w_i c)}$	(1) If.... (2) Them! (3) The Show	Generic and grammatical word sequences
<b>GPT2-Small O(LN)</b>	$\prod_i \sqrt[L]{p(w_i c, w_{<i})}$	(1) Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb (2) Star Wars: Episode VI - Return of the Jedi (3) Star Wars: Episode V - The Empire Strikes Back	Uniquely long and grammatical word sequences
<b>GPT2-Small O(LN, sum)</b>	$\prod_i p(w_i c, w_{<i})$	(1) Raiders of the Lost Ark (2) Jurassic Park (3) Ghostbusters I	Short and grammatical word sequences

Table 2: Case study of different inference methods. *Input: A user watched Raiders of the Lost Ark, Star Wars: Ep VI-Return of the Jedi, Ran, Ghostbusters II, Gandhi. Now the user may want to watch \_. Answer: Indochine*

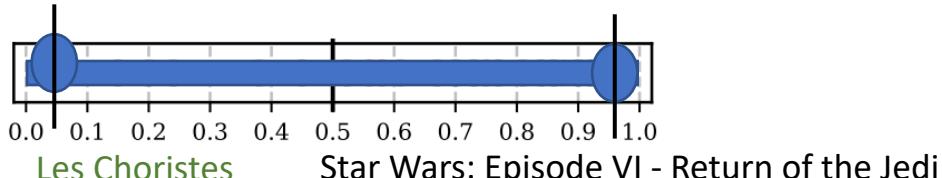
Can we calibrate the bias and do better?

# LMRecSys: Linguistic Bias Calibration

## Context Calibration

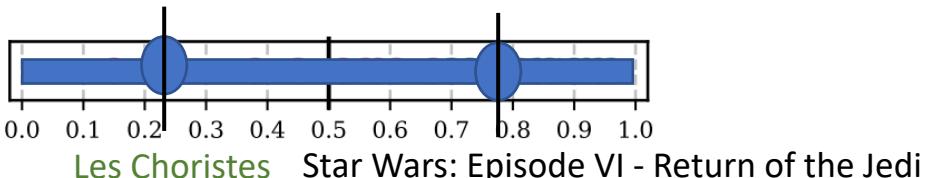
A user watched.

Now the user may want to watch \_



A user watched **A, B, C, D**.

Now the user may want to watch \_



$$s(y|x) = s'(y|x) - \alpha s(y)$$

Method	# Params	R@20
<b>Zero-shot</b>		
Random	0	0.0052
BERT-Base ItemKNN	110M	0.0599
BERT-Base O(1)	110M	0.0030
BERT-Base O(L)	110M	0.0094
BERT-Base (DAPT) O(L)	110M	0.0137
GPT2-Small O(LN)	117M	0.0667
GPT2-Medium O(LN)	345M	0.0617
GPT2-Large O(LN)	762M	0.0587
GPT2-XL O(LN)	1542M	0.0739
GPT2-Small O(LN) Weak Prompt	117M	0.0627
GPT2-Small O(LN) Strong Prompt	117M	0.0653
GPT2-Small O(LN, sum)	117M	0.0320
<b>GPT2-Small O(LN) Calibrated</b>	<b>117M</b>	<b>0.0733</b>
<b>Supervised</b>		
POP	0	0.1507
GRU4Rec	-	0.1664

Table 1: Zero-shot model performances on ML-1M.

Calibrating the linguistic bias improves the performance.

Calibrate Before Use: Improving Few-Shot Performance of Language Models. Zhao et al., 2021.

# LMRecSys: Zero-shot Recommendations

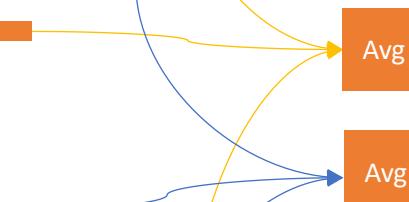
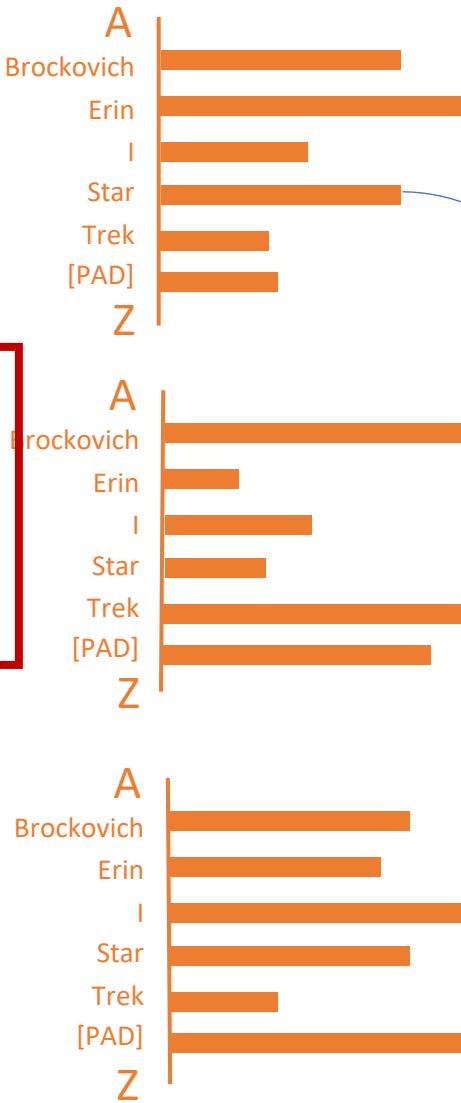
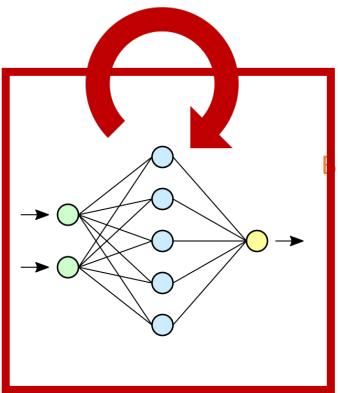
Method	# Params	R@20
<b>Zero-shot</b>		
Random	0	0.0052
BERT-Base ItemKNN	110M	0.0599
BERT-Base O(1)	110M	0.0030
BERT-Base O(L)	110M	0.0094
BERT-Base (DAPT) O(L)	110M	0.0137
GPT2-Small O(LN)	117M	0.0667
GPT2-Medium O(LN)	345M	0.0617
GPT2-Large O(LN)	762M	0.0587
GPT2-XL O(LN)	1542M	0.0739
GPT2-Small O(LN) Weak Prompt	117M	0.0627
GPT2-Small O(LN) Strong Prompt	117M	0.0653
GPT2-Small O(LN, sum)	117M	0.0320
GPT2-Small O(LN) Calibrated	117M	0.0733
<b>Supervised</b>		
POP	0	0.1507
GRU4Rec	-	0.1664

Table 1: Zero-shot model performances on ML-1M.

Model size influences the results.

# LMRecSys: Fine-tuning

A  
person  
watched  
My  
Fair  
Lady  
. Now  
they  
want  
to  
watch  
[MASK]  
[MASK]  
[MASK]



Cross-entropy Loss

$$\frac{\exp s(y|x)}{\sum_{y' \in Y_x} \exp s(y'|x)}$$

# LMRecSys: Fine-tuning

Method	MRR@20	R@20
Random	-	0.0052
POP	0.0202	0.0817
GRU4Rec	0.0302	0.0986
BERT O(1)	0.0177	0.0843
RoBERTa O(1)	0.0204	0.0843
BERT O(L)	0.0181	0.0852
BERT (DAPT) O(L)	0.0187	0.0872

Table 3: Fine-tuned model performances on ML-1M-5Star.

Fine-tuning significantly improves LMRecSys, but worse than the baseline GRU4Rec.

# LMRecSys: Domain-adaptive Pre-training (DAPT)

## 520K (150MB) movie descriptions

Her Night of Nights     Molly, a glamorous clothing model in New York, though yearning for a life of luxury, spurns the advances of her boss's son in favor of a shipping clerk, late of the backwoods.  
Rock-a-Bye Baby An average television repairman must care for the newborn triplets of his former home town sweetheart—now a famous movie star—so her career will not suffer.

Alone in the Woods: The Legend of Cambo While his parents were out working four jobs, Cambo spent his time learning how to survive in the rough backwoods of Alabama. When they went through a brutal divorce, he naturally fled to the woods to be alone. No traffic, no people, no responsibility—just pure survival. The plan was to wait out his adolescence there until he could legally live life without his parents. He ended up spending two years alone in the wild. This episode of Profiles by VICE, from director Harmony Korine, tells Cambo's story.

First Graders A documentary film about a boys school in Iran. The film shows numerous, funny and moving interviews of many different young pupils of this school summoned by their superintendent for questions of discipline. The man is not severe, but clever and fair. He teaches loyalty, fellowship and righteousness to these boys. Besides these interviews, we see scenes of this school's quotidian life

### Input:

[Description]. This is the movie [Title].

### Method:

DAPT: 50% Entire Title Mask + 50% Random Word Mask  
→ Masked Language Modeling

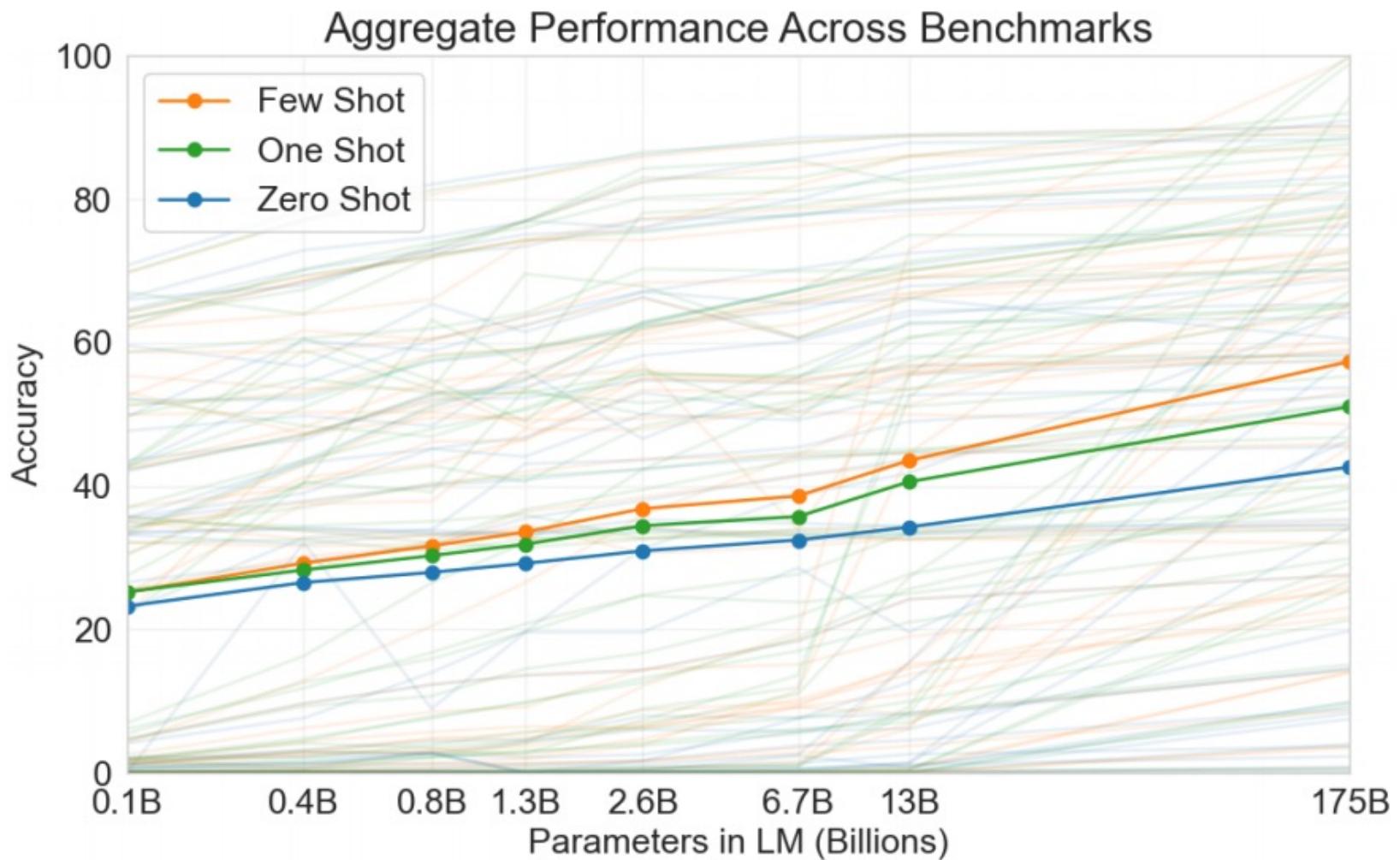
DAPT slightly improves performance.

Method	# Params	R@20
<b>Zero-shot</b>		
Random	0	0.0052
BERT-Base ItemKNN	110M	0.0599
BERT-Base O(1)	110M	0.0030
BERT-Base O(L)	110M	0.0094
BERT-Base (DAPT) O(L)	110M	0.0137

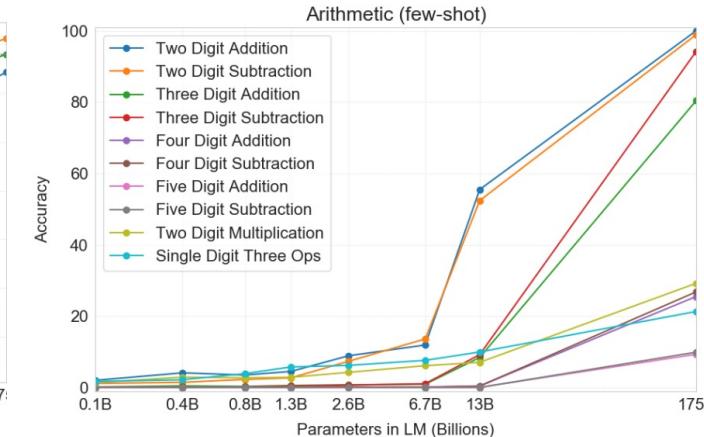
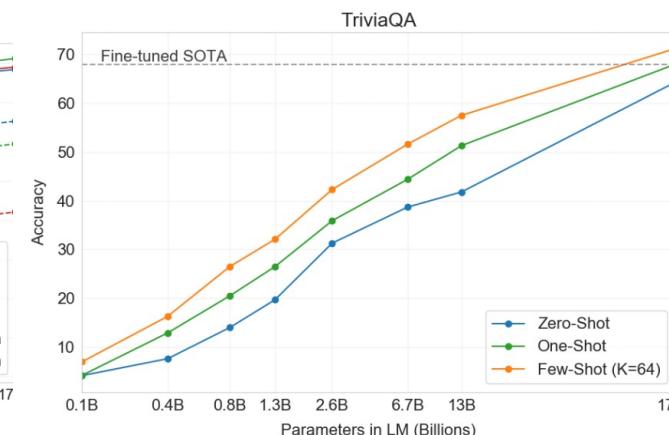
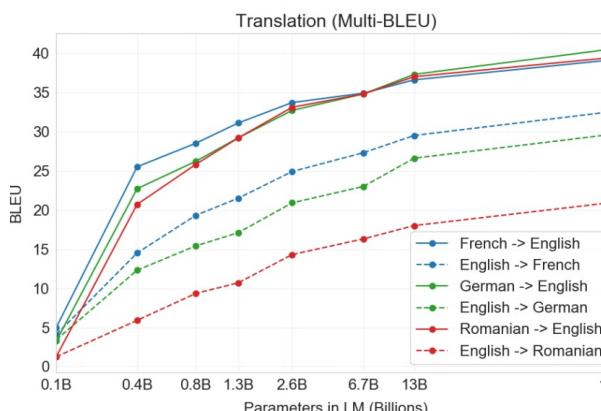
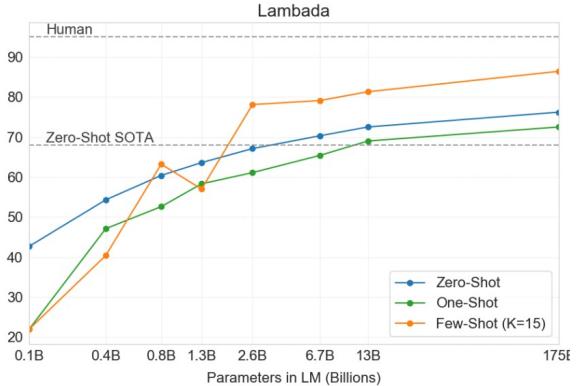
Method	MRR@20	R@20
Random	-	0.0052
POP	0.0202	0.0817
GRU4Rec	0.0302	0.0986
BERT O(1)	0.0177	0.0843
RoBERTa O(1)	0.0204	0.0843
BERT O(L)	0.0181	0.0852
BERT (DAPT) O(L)	0.0187	0.0872

Table 3: Fine-tuned model performances on ML-1M-5Star.

# Scaling Law



# Scaling Law



## Perception

### (System 1)

## Reasoning

## – (System 2)



## Recommendation

# Case Study

Text Generation

The Truman Show is a movie about

Compute

Computation time on cpu: 10.5696 s

The Truman Show is a movie about a man who lives in isolation, alone in a house in Nashville. He is the only person in attendance as his wife leaves to be with a friend. He receives a phone call that sends him and his wife on

Text Generation

The Truman Show is a movie about

Compute

Computation time on cpu: 10.4856 s

The Truman Show is a movie about a man who moves into a small community and experiences a series of problems that gradually become bigger and more severe until his life comes to an end. The movie takes place in a town and has a character in

Text Generation

The Truman Show is a movie about

Compute

Computation time on cpu: 10.5156 s

The Truman Show is a movie about the life of a small town middle-aged couple. Both of them are played by Harrison Ford and Meryl Streep. Truman is an accountant who is constantly struggling to make ends meet and get by, and M

Text Generation

The Truman Show is a movie about

Compute

Computation time on cpu: 10.04399999999999 s

The Truman Show is a movie about a man whose world collapses the moment he is unable to find an audience for his act, at a hotel. As the actors and the audiences continue to drift away, it is up to Jeff Daniels, the guy who

# Case Study

Canvas

Quickstart

 Clear all  Share

The movie, The Truman Show,

, starring Jim Carrey, is a 1998 American satirical science fiction film directed by Peter Weir. The screenplay by Andrew Niccol was adapted from Niccol's 1997 novel of the same name. The film tells the story of Truman Burbank, a man who is unwittingly placed in a televised reality show that broadcasts every aspect of his life without his knowledge.

## Plot

Truman Burbank is a man whose life is unknowingly broadcast around the clock by a vast media conglomerate. His entire life has been meticulously planned, controlled, and manipulated by Christof, the show's creator and executive producer. Christof orchestrates Truman's entire life, from his marriage to Meryl, to his job at the Seahaven Aquarium, to his friends, and even his romance with Sylvia.

Truman's entire life takes place within Seahaven, a fabricated community set

Configuration

Model

j1-jumbo (178B) 

Max completion length 135 

1 2048

Temperature 0 

0 1

Top P 0.5 

0 1

Stop sequences

Type string and press Tab

Alternative tokens 

# Case Study

Canvas [Quickstart](#)

[Clear all](#) [Share](#)

A user watched Jaws, Saving Private Ryan, The Good, the Bad and the Ugly, Run Lola Run, Goldfinger. Now the user may want to watch **something funny and light-hearted comfort him after having seen some horrors.**

**A user likes the TV show Big Bang Theory, and decides to watch the full series (200 episodes) with his DVR, in a chronological order.**

## Other similar problems

With these guiding principles in place, there are a few notable tasks that the scheduling engine at NLNet Labs can identify.

For example, a user may like

Movie posters for Jaws, Saving Private Ryan, and Run Lola Run are displayed. The Jaws poster shows a shark attack scene. The Saving Private Ryan poster shows soldiers in a field. The Run Lola Run poster shows a woman running.

### Jaws

PG · 1975 · Thriller/Adventure · 2h 10m

### Saving Private Ryan

R · 1998 · War/Action · 2h 50m

### Run Lola Run

R · 1998 · Thriller/Action · 1h 22m

## Goldfinger / Genres

Action

Spy

Adventure

Thriller

## Configuration

### Model

j1-jumbo (178B)

Max completion length 64  
1 2048

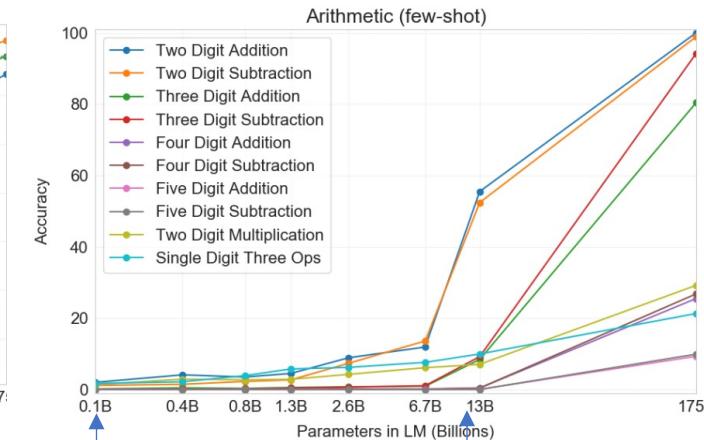
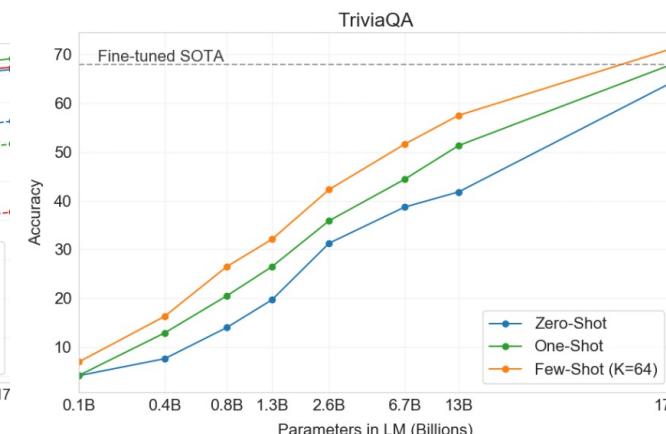
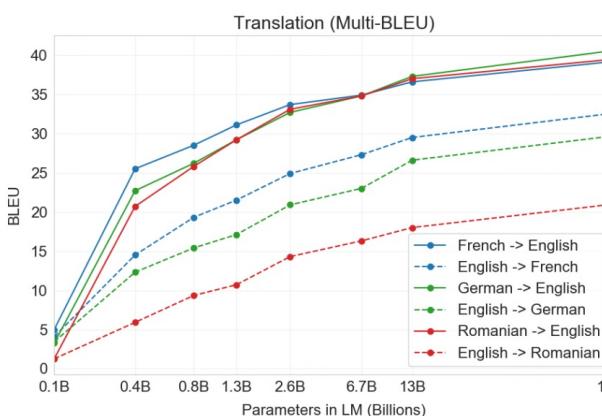
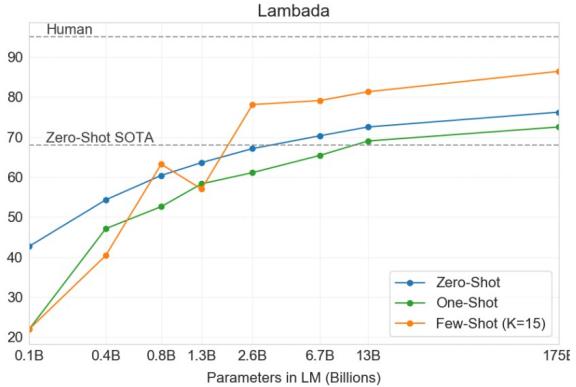
Temperature 1  
0 1

Top P 1  
0 1

Stop sequences  
Type string and press Tab

Alternative tokens

# Scaling Law



## Perception

(System 1)

## Reasoning

# GPU trainable (we are here)

Largest publicly available  
(tried zero-shot here)



## Recommendation

## Overview

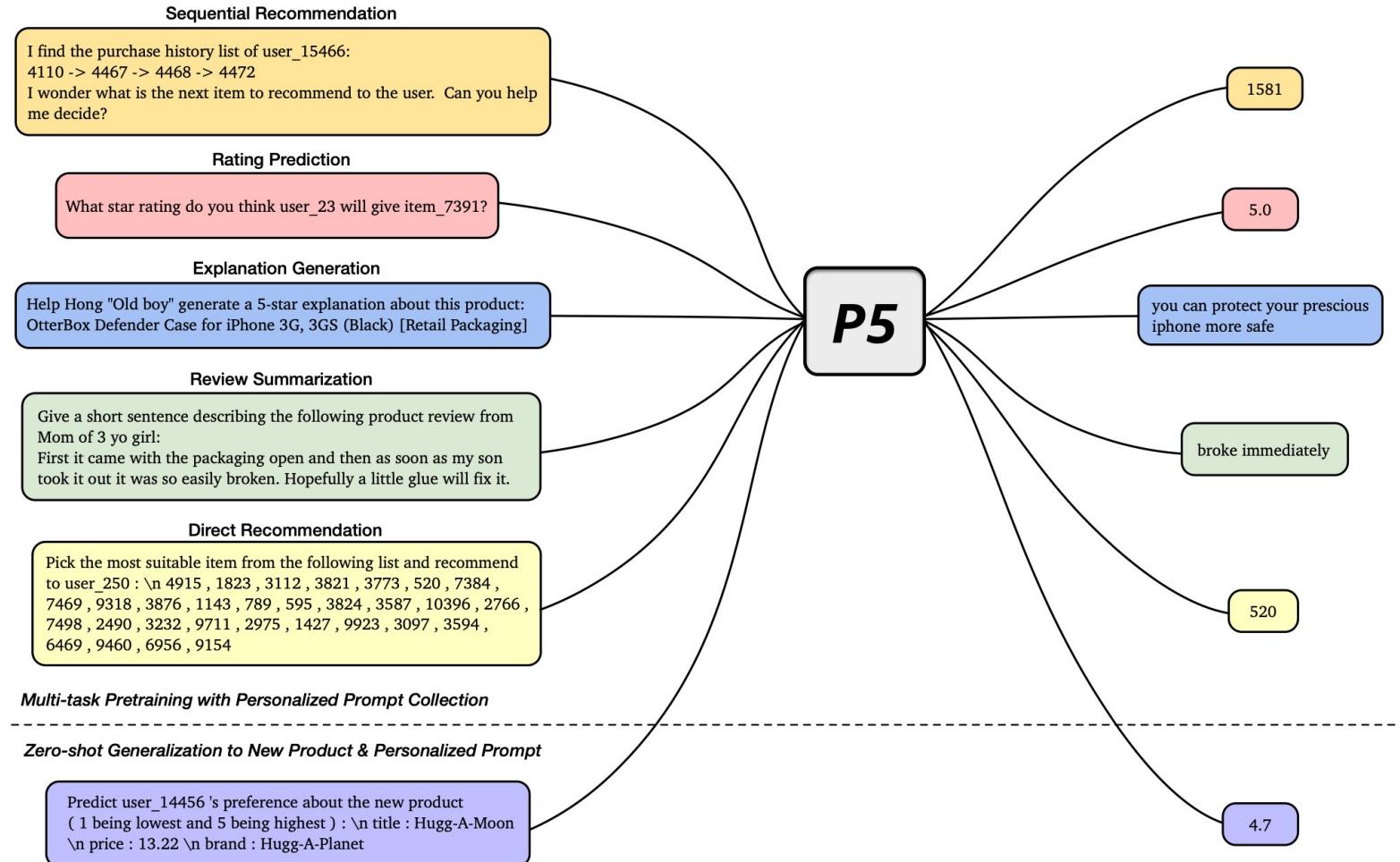
- Recent Advances in Large Language Models
- Language Models as Recommender Systems
- Follow-up Works

# Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt & Predict Paradigm (P5)

Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, Yongfeng Zhang

Department of Computer Science, Rutgers University, NJ 08854, US

{sg1309,shuchang.syt.liu,zuohui.fu,yingqiang.ge,yongfeng.zhang}@rutgers.edu

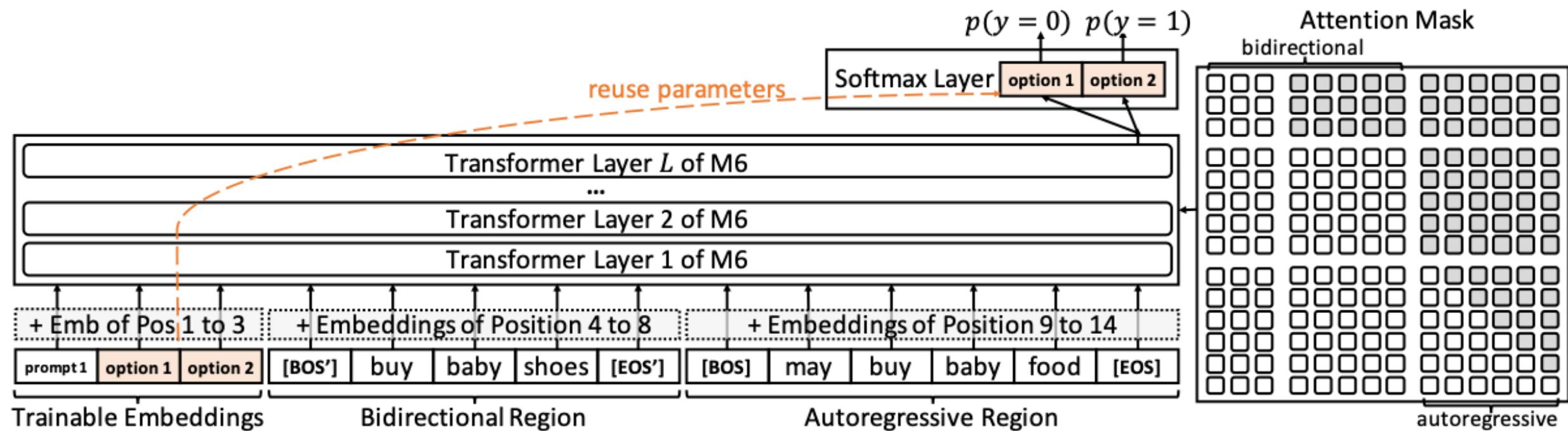


# M6-Rec: Generative Pretrained Language Models are Open-Ended Recommender Systems

Zeyu Cui\*, Jianxin Ma\*, Chang Zhou, Jingren Zhou, Hongxia Yang

{zeyu.czy,jason.mjx,ericzhou.zc,jingren.zhou,yang.yhx}@alibaba-inc.com

DAMO Academy, Alibaba Group



# PROMPTCAST: A NEW PROMPT-BASED LEARNING PARADIGM FOR TIME SERIES FORECASTING

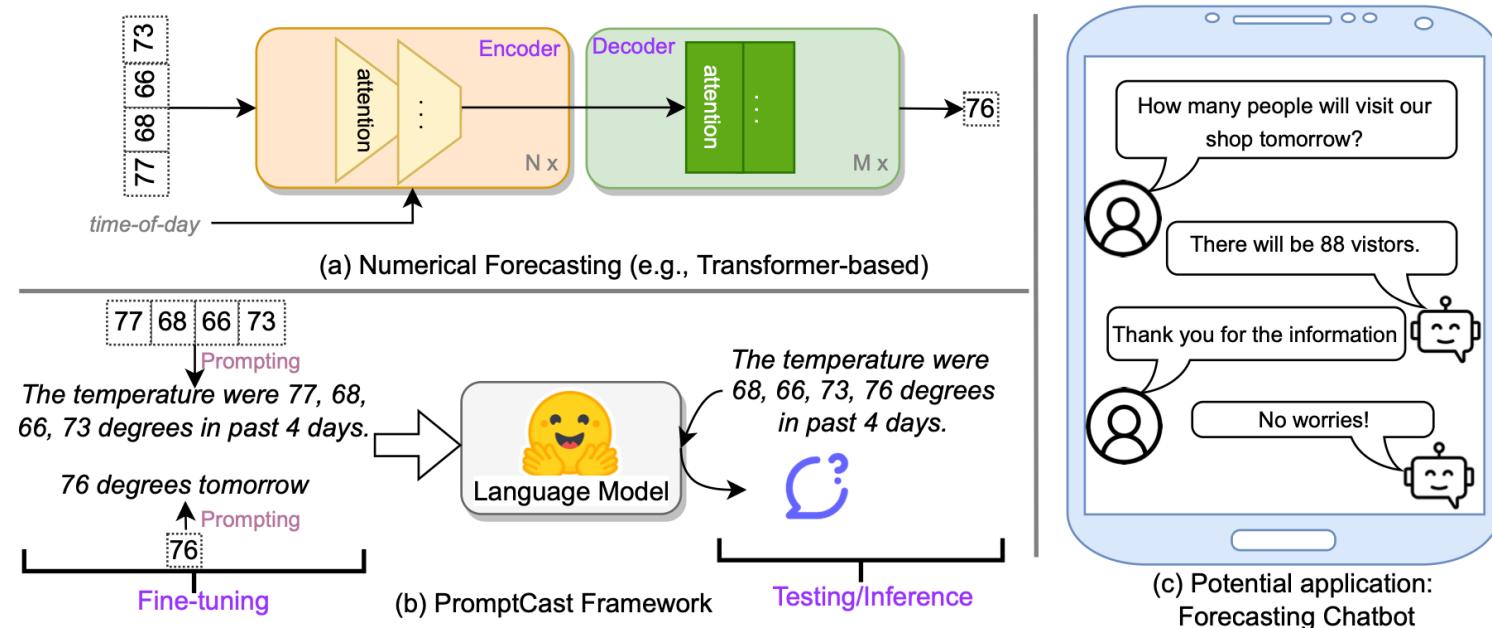


Figure 1: Conceptual illustrations of (a) existing numerical-based forecasting; (b) the framework of the proposed PromptCast; (c) a potential forecasting chatbot application based on PromptCast.



# LIFT Language-Interfaced Fine-Tuning for Non-Language Machine Learning Tasks

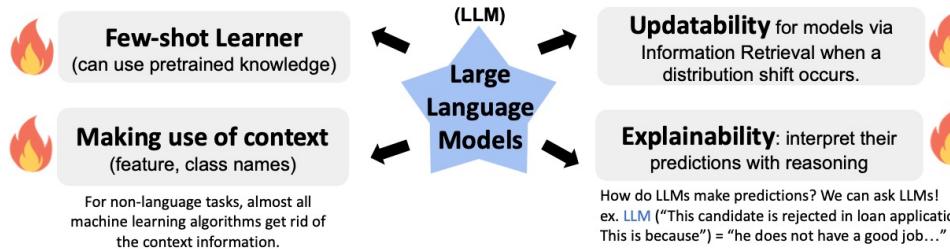
Tuan Dinh\*, Yuchen Zeng\*, Ruisu Zhang, Ziqian Lin, Michael Gira, Shashank Rajput, Jy-yong Sohn, Dimitris Papailiopoulos, Kangwook Lee  
University of Wisconsin-Madison



- ❖ **Objective:** Leverage abilities of large pretrained language models to better solve non-language tasks.
- ❖ **Our Idea:** Convert everything into sentences and finetune a pretrained language model!
- ❖ **Findings:** (1) LIFT performs comparably well on a suite of tasks: classification (e.g., tabular data) and regression tasks. (2) LIFT is highly robust to outliers. (3) LIFT can be improved by appropriate prompting, two-stage fine-tuning, data augmentation. (4) LIFT can be used for data generation, in-context learning.

## 1. Language Models for Non-Language Tasks

Non-language tasks include tabular classification, and regression.

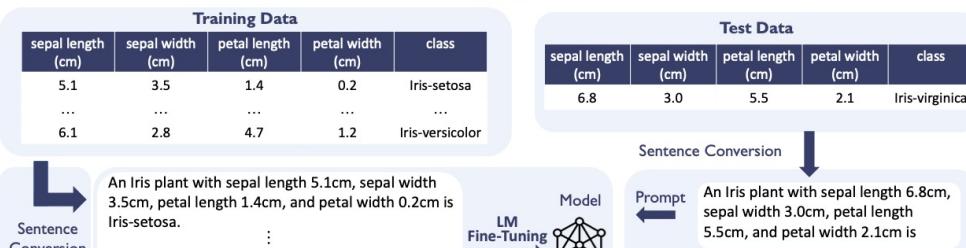


### Key Challenge

Can we use Large Language Models for non-language tasks?

## 2. Language-Interfaced Fine-Tuning (LIFT)

x: non-language data → x: sentence format → LLM → y: sentence format → y: non-language label



### LIFT Training

**Acknowledgement.** This work was supported by NSF Award DMS-2023239, the Understanding and Reducing Inequalities Initiative of the University of Wisconsin-Madison, and the Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation.

- References**
- Kevin Lu, et al. (2021). Pretrained transformers as universal computation engines.

## 3. Findings

### LIFT Can Perform Classification

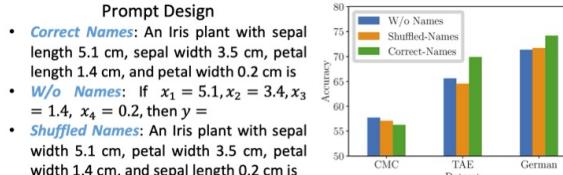
More than 20 classification tasks on synthetic, tabular, and image data...

Dataset	LogReg	XGBoost	LIFT/GPT-J	LIFT/GPT-3
Synthetic Data				
two circles	49.83±4.18	79.25±0.35	75.92±1.65	<b>81.42±0.82</b>
blobs	<b>96.75±0.00</b>	96.17±0.12	96.17±0.59	96.67±0.24
moons	88.58±0.12	99.83±0.12	99.58±0.42	<b>100.00±0.00</b>
Tabular Data (OpenML)				
Hill-Valley	77.78±0.00	59.26±0.00	<b>100.00±0.00</b>	<b>99.73±0.19</b>
IRIS	96.67±0.00	<b>100.00±0.00</b>	96.67±0.00	97.0±0.00
TAE	45.16±4.56	<b>66.67±8.05</b>	61.29±6.97	65.59±6.63
Wine	<b>100.00±0.00</b>	97.22±0.00	93.52±1.31	92.59±1.31
Image Data				
MNIST	91.95±0.69	97.69±0.04	97.01±1.15	<b>98.15±0.67</b>
Fashion MNIST	85.59±0.09	<b>90.19±0.04</b>	85.10±0.19	90.18±0.12

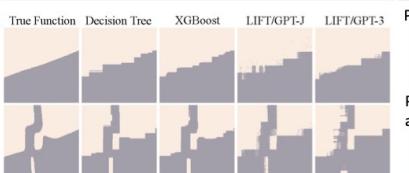
### LIFT Can Make Use of Feature Names

#### Prompt Design

- Correct Names:** An Iris plant with sepal length 5.1 cm, sepal width 3.5 cm, petal length 1.4 cm, and petal width 0.2 cm is
- W/o Names:** If  $x_1 = 5.1, x_2 = 3.4, x_3 = 1.4, x_4 = 0.2$ , then  $y =$
- Shuffled Names:** An Iris plant with sepal width 5.1 cm, petal width 3.5 cm, petal length 1.4 cm, and sepal length 0.2 cm is



### Visualization of Decision Boundaries



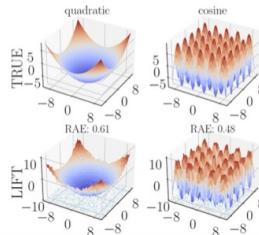
### Fine-Tuning v.s. In-Context Learning

**Remark:** when the target tasks requires fewer training samples , one can replace fine-tuning with incontext-learning in our language-interfaced procedure.

Dataset	#Prompts	MCC	In-Context	LIFT/Subset	LIFT/Full-data	In-Context	LIFT/Subset	LIFT/Full-data	GPT-3
Breast	35	70.69	56.90±19.51	<b>58.62±2.44</b>	64.94±11.97	62.07±1.41	<b>70.69±0.00</b>	71.26±1.62	
TAE	50	35.48	<b>34.33±1.47</b>	32.26±9.50	61.29±4.56	37.64±4.02	33.33±1.52	65.59±6.63	
Customers	29	68.18	56.06±17.14	<b>59.85±2.84</b>	85.23±1.61	60.61±1.42	<b>63.26±6.96</b>	84.85±1.42	

\*MCC denotes the majority class classifier.

### LIFT Can Approximate Functions



## 4. Future Directions

- Many ways to make LIFT even better
- Language models are getting better and better!
- GPT4 is coming soon!
- Different prompting, chain of thought (CoT), etc.
- ... Let's think step by step."
- [In progress] Using language description but with customized layer or loss function.
- Something between LIFT and Frozen TF[1].

Thanks! Questions?