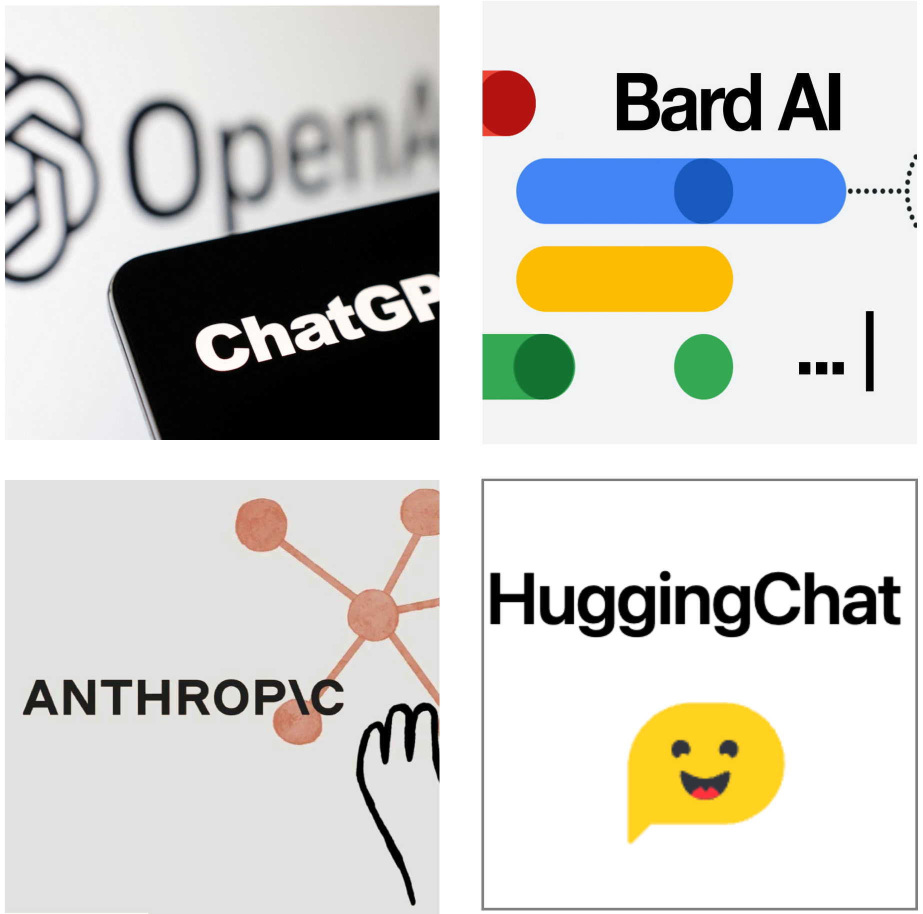
Large Language Models: Revisiting Few Mysteries

Daniel Khashabi

JOHNS HOPKINS UNIVERSITY

Please don’t hesitate to
stop me and ask questions.

The success we dreamed of



Language models that are remarkably capable at solving many important NLP benchmarks.

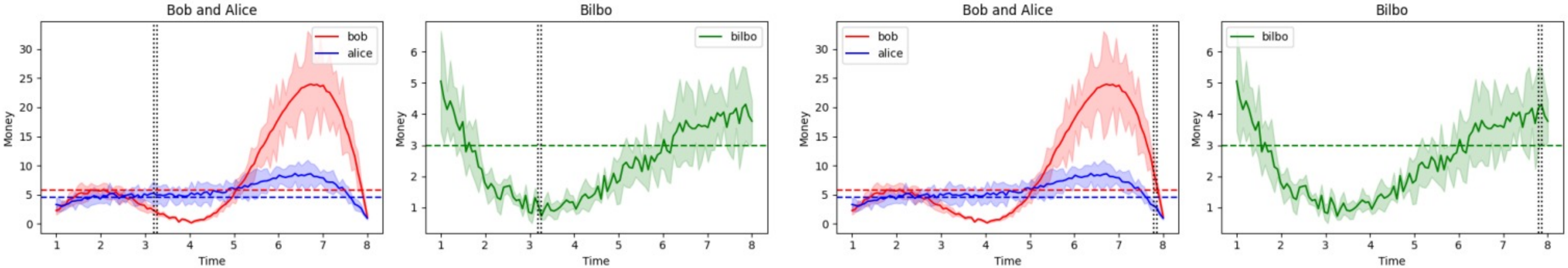
[Bubeck et al. 2023]

The success we dreamed of

Prompt: Can you generate a pyplot for the following data: $x = [1, 3, 5, 6, 8], y1 = [2, 3, 5, 18, 1], y2 = [3, 5, 6, 8, 1], y3 = [5, 1, 2, 3, 4], y4 = [9, 7, 2, 3, 1]$. I want $y1, y2$ to be in the same plot, but $y3$ is in another plot next to that plot, $y4$ is in below. I want the legend of $y1$ to be "bob", $y2$ to be "alice", $y3$ to be "bilbo", $y4$ to be "allie". I want the x-axis to be labeled with "time" and y axis to be labeled with "money". I want to add a 10 %-40% random error bar to each curve, through all times (including non-integers). I want smoothed curves to show the plot, and smoothed error bar. Do not use linear interpolation, use smooth interpolation! I want to also add some small zig-zag to the smoothed curve to make it look more real. I want to put a baseline as the mean of each line. I want to put a pie chart below indicating the fraction of the four people in each time step. I also want an animation to show how the fractions are changing in the pie chart in **continuous time**. Interpolate the missing fractions! I also want vertical line animation in other three plots to match the pie chart. I want the fanciest plot. Please add as many fancy things as possible.

Here is a possible code to generate a pyplot for the following data, with some comments to explain the steps:

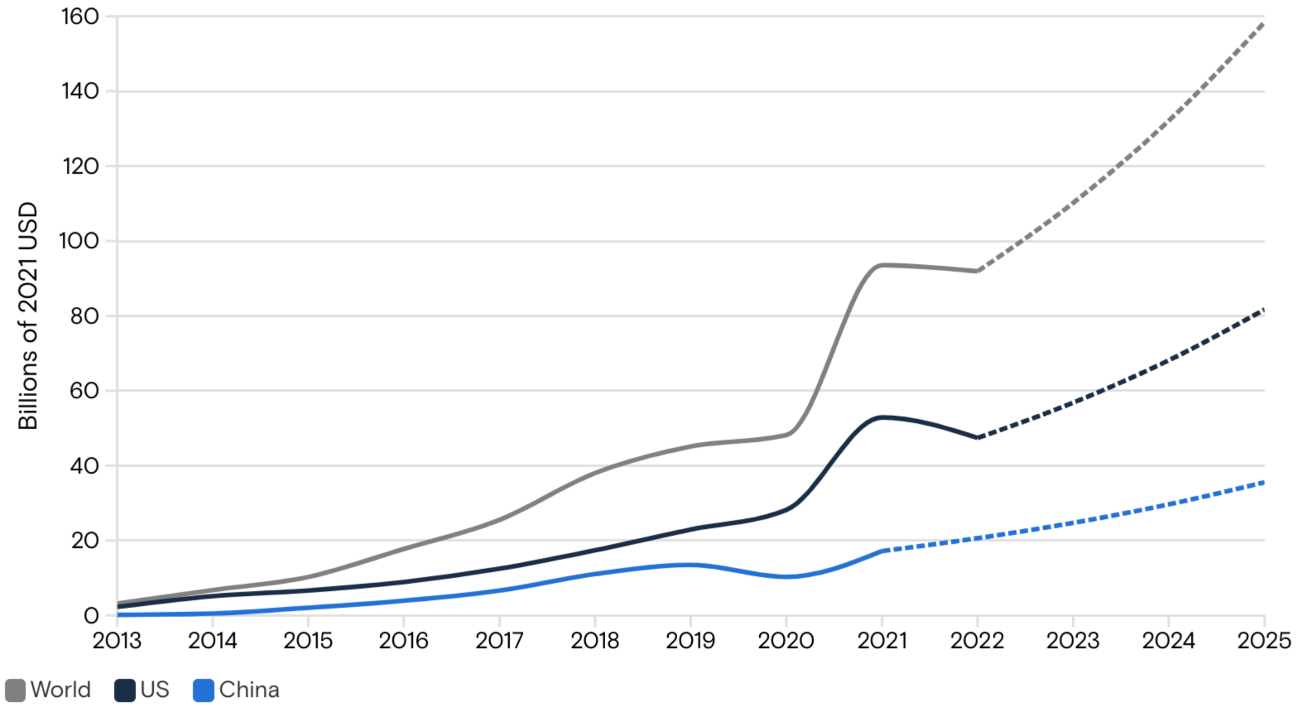
GPT-4: [We give two snapshots from the animation resulting from the produced code]



Accelerated Industrialization of AI

AI investment is likely to grow in the next three years

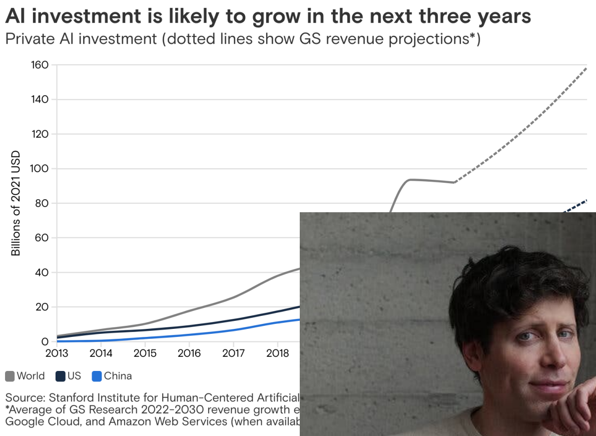
Private AI investment (dotted lines show GS revenue projections*)



Source: Stanford Institute for Human-Centered Artificial Intelligence, Goldman Sachs Research • *Average of GS Research 2022-2030 revenue growth estimates for Microsoft Azure, NVIDIA, Google Cloud, and Amazon Web Services (when available)

Goldman Sachs

Accelerated Industrialization of AI



Accelerated industrialization of AI based on market competition entails diverging missions.

Remarkable progress
but many questions remain open.

- Questions about
  - optimality of architectures,
  - limits of their controllability,
  - scope of machine innovations,
  - effective interaction with humans, … .

- **Today:** Revisit two interrelated technological pieces that deserve further deliberation.

Today

- Revisiting …

In-Context Learning

Alignment of chatbots

Today

- Revisiting …

In-Context Learning

Alignment of chatbots

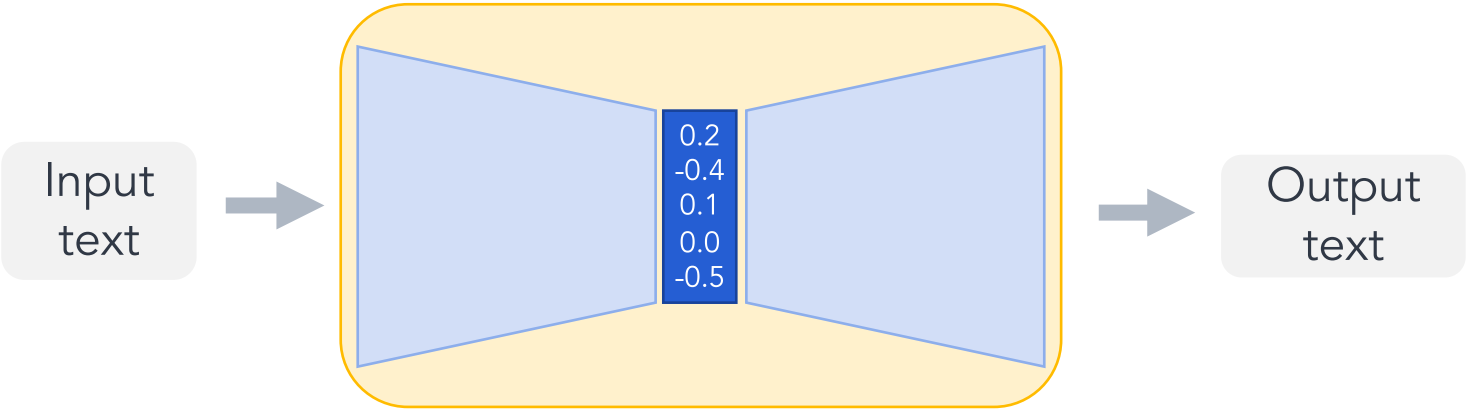
Background

Language Models

[Bengio et al. '04, Peters et al. '18, Raffel et al. '20, Brown et al. '20, many others]

Background

Language Models



[Bengio et al. ’04, Peters et al. ’18, Raffel et al. ’20, Brown et al. ’20, many others]

Background

Language Models

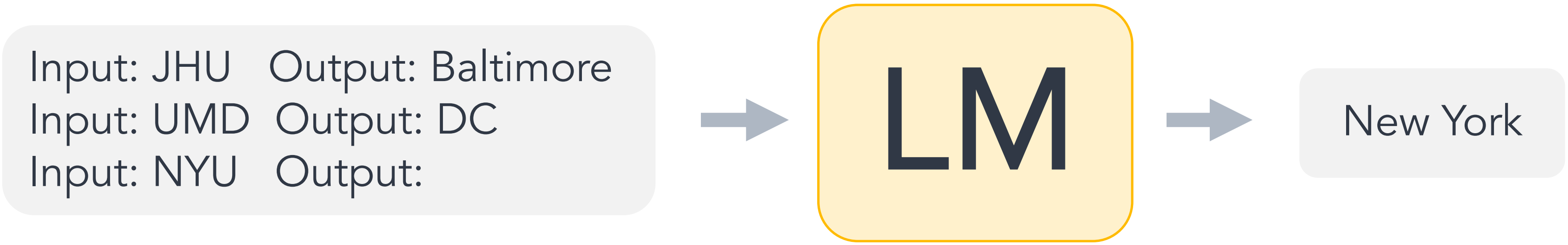
Johns Hopkins University is in ________. → LM → Baltimore

Simple facts

[Bengio et al. ’04, Peters et al. ’18, Raffel et al. ’20, Brown et al. ’20, many others]

Background

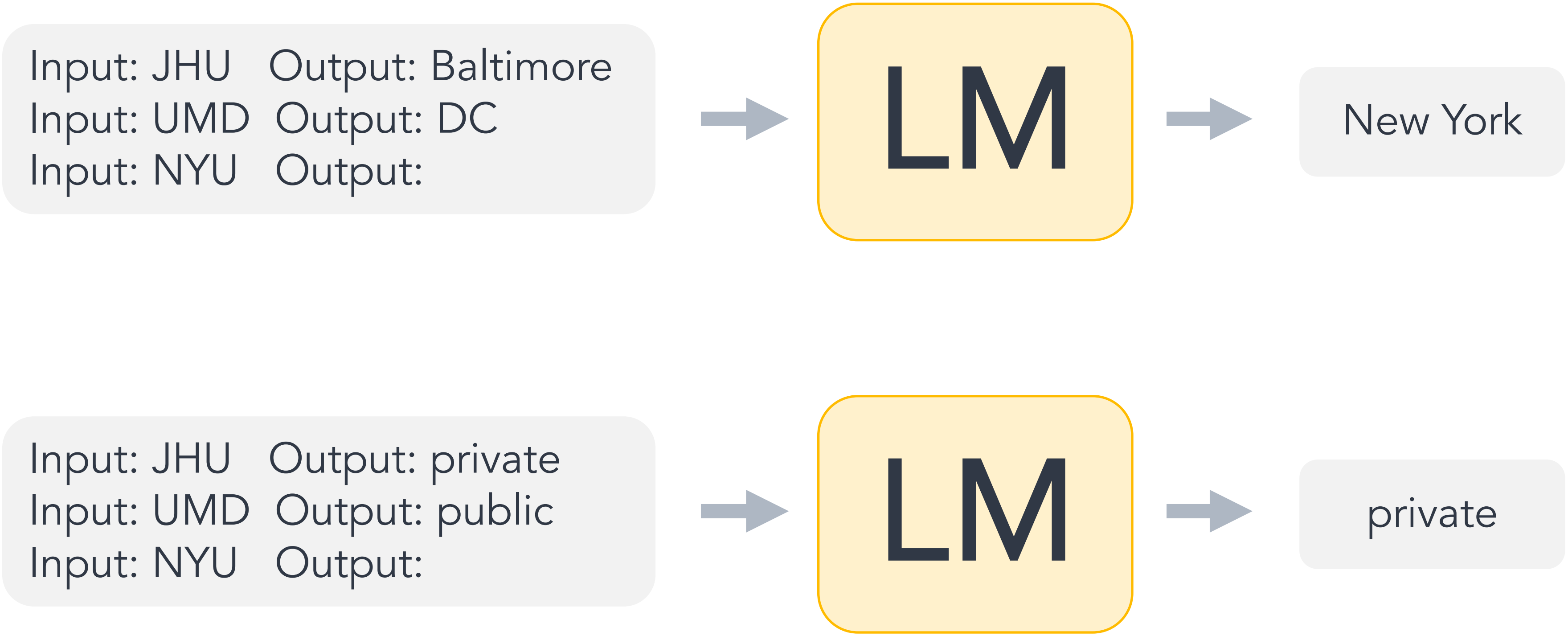
In-context learning emerges from pre-training



[Brown et al., 2020]

Background

In-context learning emerges from pre-training

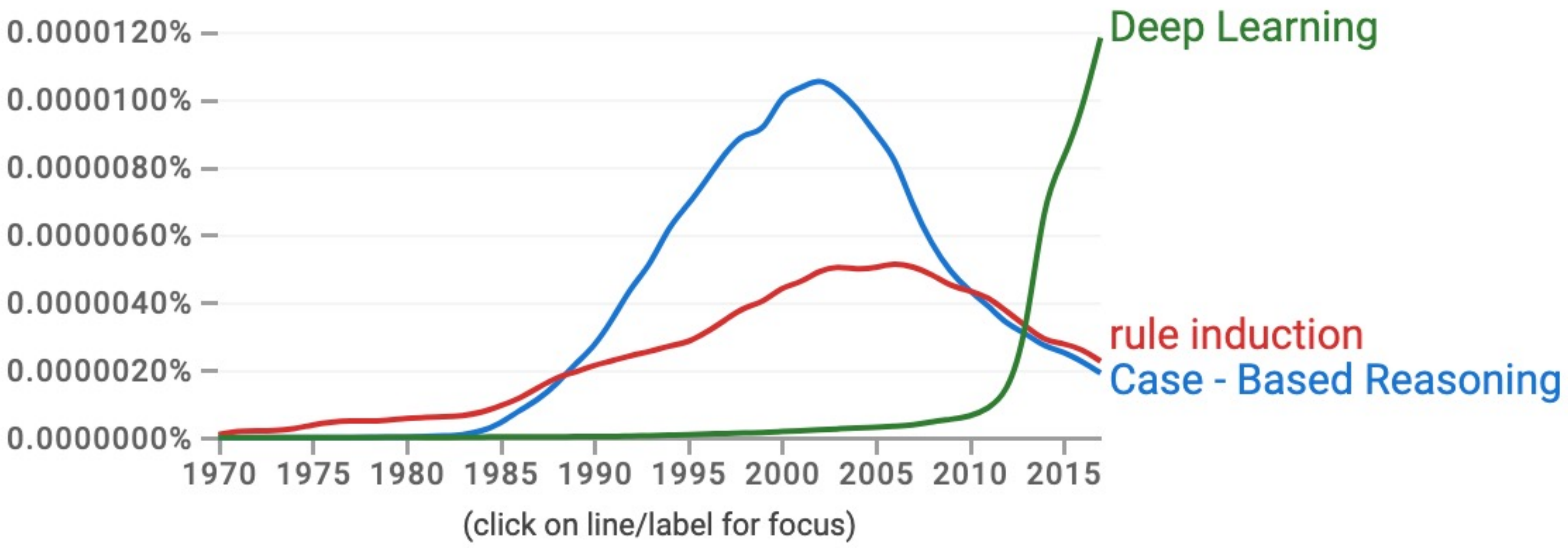


[Brown et al., 2020]

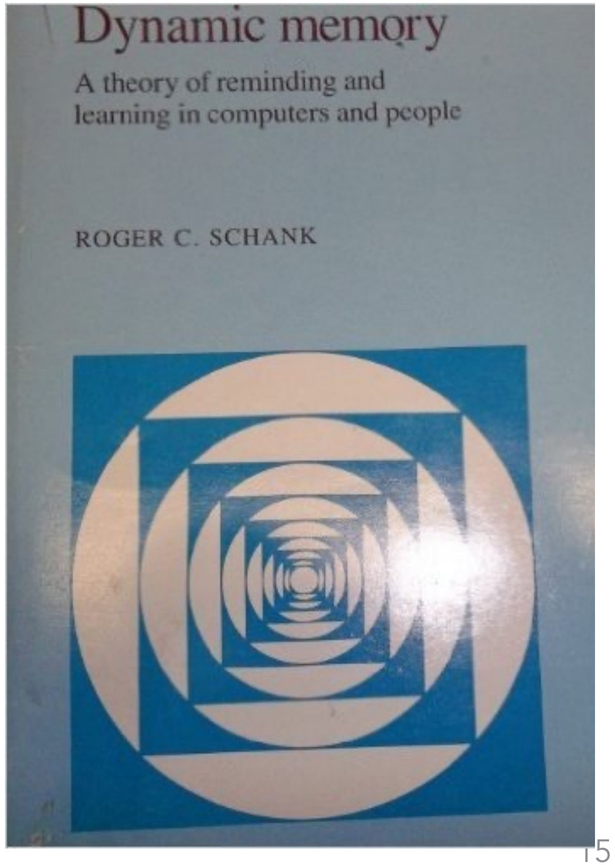
Background

This is an old dream come true!

Case-based reasoning, rule-induction, dynamic memory, analogical reasoning, …



[Google n-grams]



In-context learning: well-studied yet elusive.

- What we understand:
  - ICL improves with scale. [Brown et al. 2020; Srivastava et al. 2023]
  - ICL is brittle. [Min et al., 2022; Mishra et al., 2022]
  - ICL as a probabilistic inference. [Muller et al. 2021; Xie et al. 2021]
- Still no framework that fully explains and predicts its nuts and bolts.

Explaining ICL via Gradient Descent

- Is it possible that ICL is secr xecuting GD during inference?
- We have known GD for a l e.

Transformers Learn In-Context b

Johannes von Oswald [1 2] Eyvind Niklas
Alexander Mordvintsev [2] An

ICM

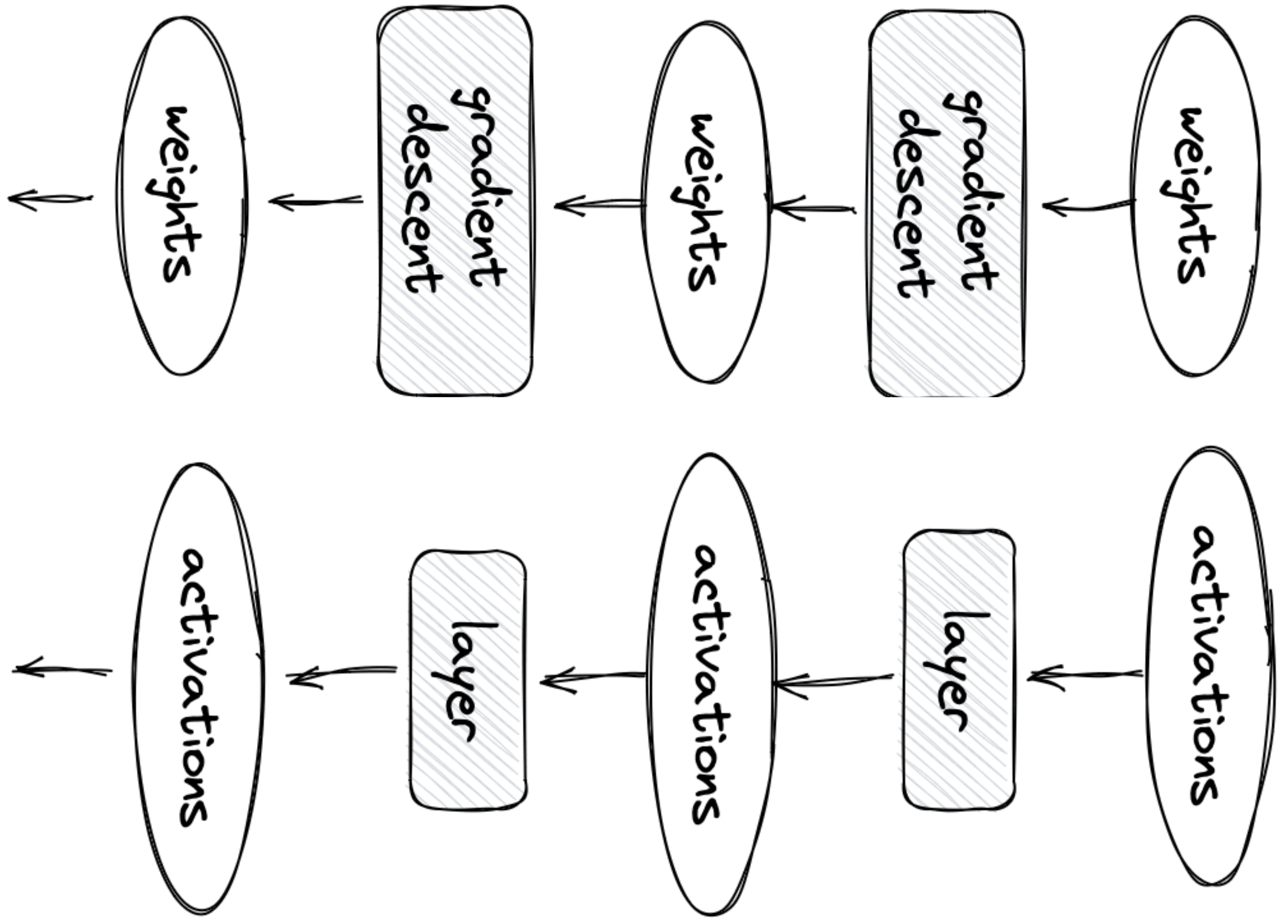
RITHM IS IN-CONTEXT LEARN-
WITH LINEAR MODELS

Ekin Akyürek [1,2,a] Dale Schuurmans [1] Jacob Andreas [*2] Tengyu Ma [*1,3,b] Denny Zhou [*1]

ICLR 2023

Dai et al. 2022; Garg et al. 2022; Zhang et al. 2023;
Ahn et al. 2023; Raventos et al. 2023; Li et al. 2023; Guo et al. 2023; …

Basic idea: gradient computation in forward process



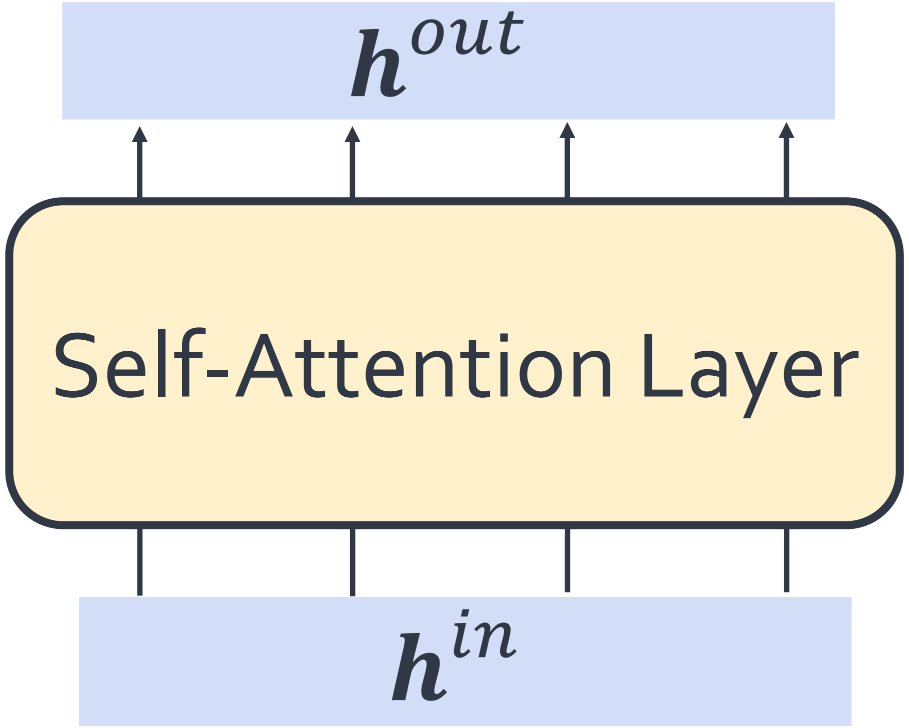
Gradient descent

Forward propagation

(photo credit: Blaine on lesswrong)

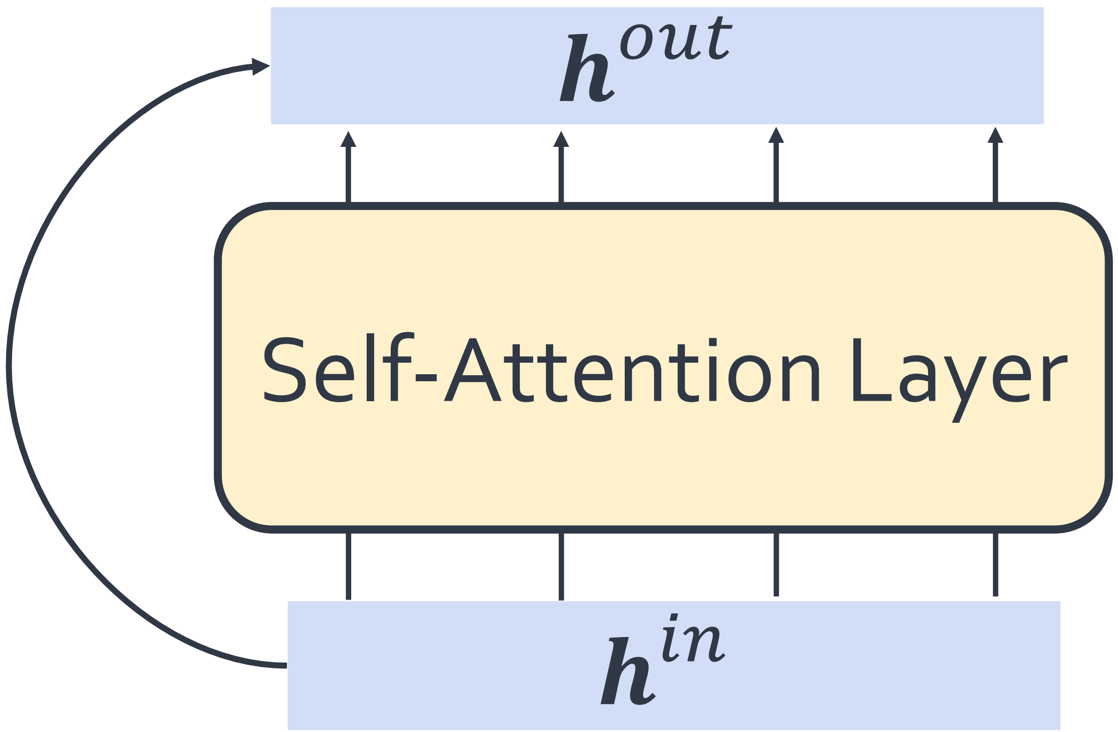
A Self-Attention (SA) Layer

$$\boldsymbol{h}^{out} = \mathrm{SA}(\boldsymbol{h}^{in};\ \boldsymbol{W}_q, \boldsymbol{W}_k, \boldsymbol{W}_v)$$



A Self-Attention (SA) Layer

$$\boldsymbol{h}^{out} = \boldsymbol{h}^{in} + \mathrm{SA}(\boldsymbol{h}^{in}; \boldsymbol{W}_q, \boldsymbol{W}_k, \boldsymbol{W}_v)$$



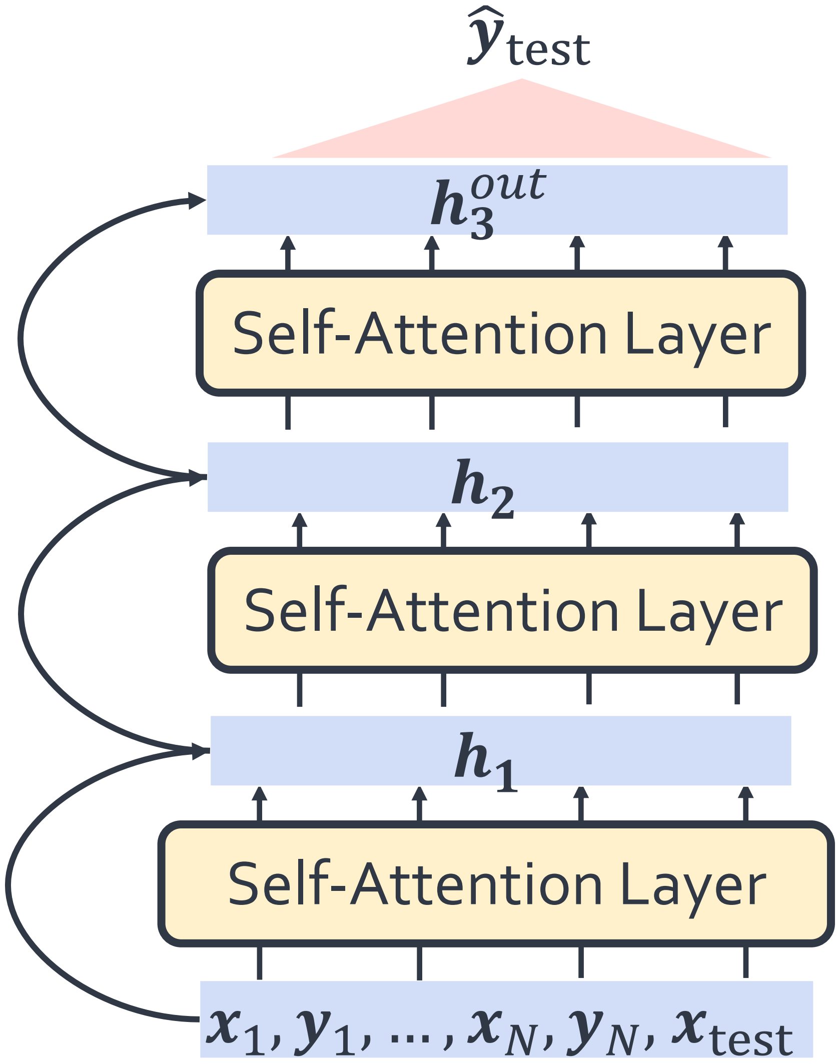
A SA Layer vs. a GD update

$$\boldsymbol{h}^{out} = \boldsymbol{h}^{in} + \underbrace{\mathrm{SA}(\boldsymbol{h}^{in}; \boldsymbol{W}_q, \boldsymbol{W}_k, \boldsymbol{W}_v)}$$

Each layer simulate an implicit gradient update?

$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^{t} \; \overbrace{- \eta \times \nabla \mathcal{L}}$$

Results: Transformers can != does implement GD



demonstrations

$$\hat{y}_{\text{test}} = f(\boldsymbol{x}_{\text{test}}; \boldsymbol{w}^3)$$

$$\boldsymbol{w}^3 = \boldsymbol{w}^2 - \frac{\eta}{N}\sum_{i=1}^{N} \nabla\ell(f(\boldsymbol{x}_i), \boldsymbol{y}_i)$$

$$\boldsymbol{w}^2 = \boldsymbol{w}^1 - \frac{\eta}{N}\sum_{i=1}^{N} \nabla\ell(f(\boldsymbol{x}_i), \boldsymbol{y}_i)$$

$$\boldsymbol{w}^1 = \boldsymbol{w}^0 - \frac{\eta}{N}\underbrace{\sum_{i=1}^{N} \nabla\ell(f(\boldsymbol{x}_i), \boldsymbol{y}_i)}_{\text{gradients of the empirical loss on the demonstrations}}$$

$$\boldsymbol{w}^0$$

[Von Oswald et al. 2023; others]

!= does

Results: Transformers can implement GD

Theorem [von Oswald et al., among others]: There exists self-attention weights that, ICL simulates GD, for a fixed well-defined task family.

How strong of a claim are we making here?
Do they hold in real practice?

What existing work shows:

∃

Theorem [von Oswald et al., among others]: There exists self-attention weights that, ICL simulates GD, for a fixed well-defined task family.

Do the existing results generalize to realistic settings?

What is more interesting and realistic:

∀

Hypothesis [ICL≈GD hypothesis]: For <u>any</u> pre-trained Transformer weights , ICL is equivalent to GD, for <u>any</u> well-defined task family.

Do Pretrained Transformers Really Learn In-Context by Gradient Descent?

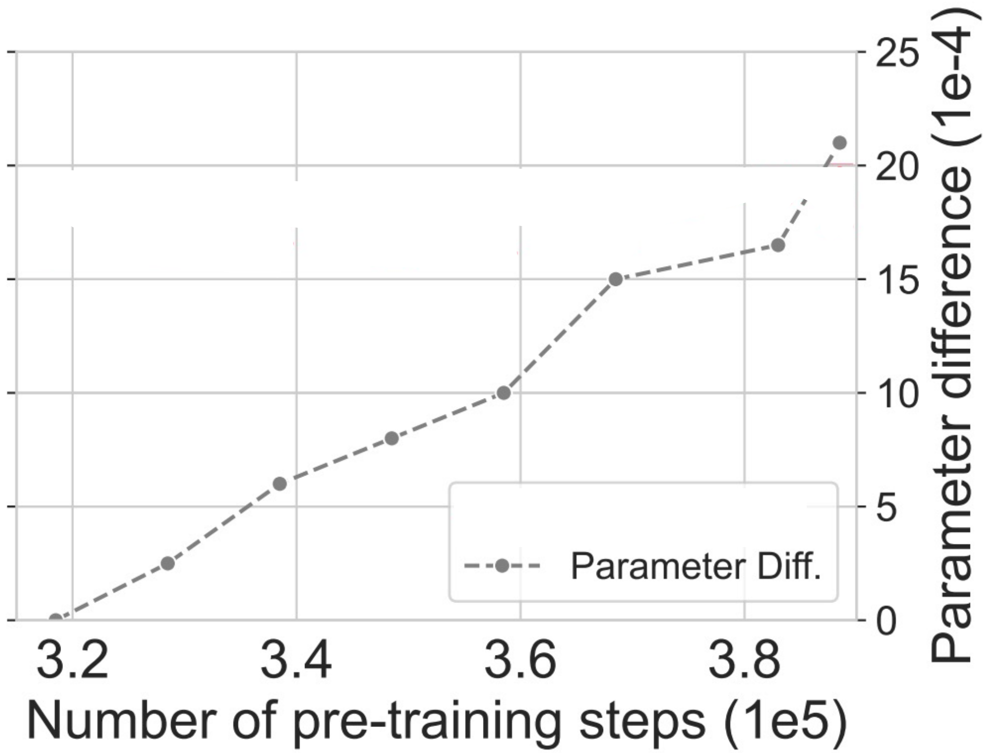
Lingfeng Shen, Aayush Mishra, Daniel Khashabi



https://arxiv.org/abs/2310.08540

How realistic is it to proves ICL≈GD for fixed weights?

- GPT-J's ICL ability does not change much over time during training, while the parameters change steadily.
- There are many ICL-inducing parameters.



Therefore, to prove **ICL≈GD hypothesis**, showing it for a single choice of parameters is not enough.

ICL vs GD: End task comparison

Possible values:
`World, Sports, Business, Tech`

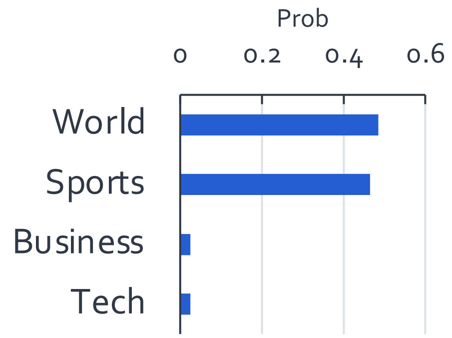
```
Input: Rookie Taylor Wins Playoff at Tahoe
Output:
```

→ LM → 

demonstrations

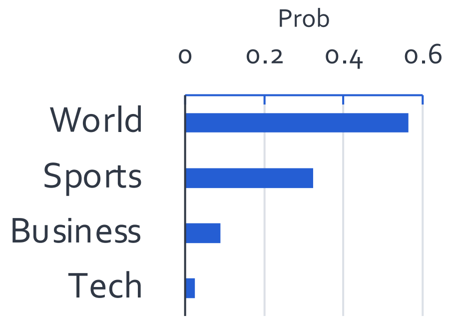
```
Input: Apple recalls 15in PowerBook batteries
Output: Tech
Input: Major attack by rebels on Nepalese town
Output: World
...
Input: Rookie Taylor Wins Playoff at Reno-Tahoe
Output:
```

→ In-context LM →




```
Input: Rookie Taylor Wins Playoff at Tahoe
Output:
```

→ Gradient-descent LM →



ICL vs GD: End task comparison

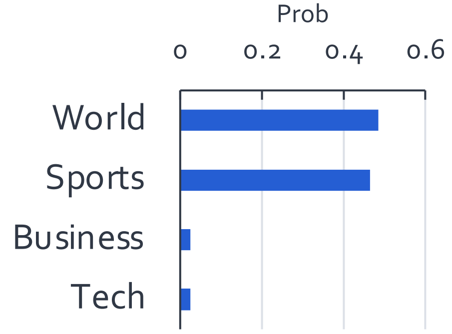
demonstrations

```
Input: Apple recalls 15in PowerBook batteries
Output: Tech
Input: Major attack by rebels on Nepalese town
Output: World
...
Input: Rookie Taylor Wins Playoff at Reno-Tahoe
Output:
```

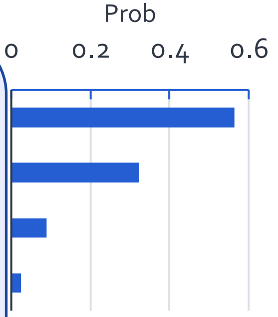
In-context

LM

Possible values:
`World, Sports, Business, Tech`




```
Input:
Output
```



Hypothesis: If two adaptation algorithms consistently lead to the same distribution on any tasks, they must be equivalent.

ICL vs GD: End task comparison

demonstrations

```
Input: Apple recalls 15in PowerBook batteries
Output: Tech
Input: Major attack by rebels on Nepalese town
Output: World
...
Input: Rookie Taylor Wins Playoff at Reno-Tahoe
Output:
```

In-context

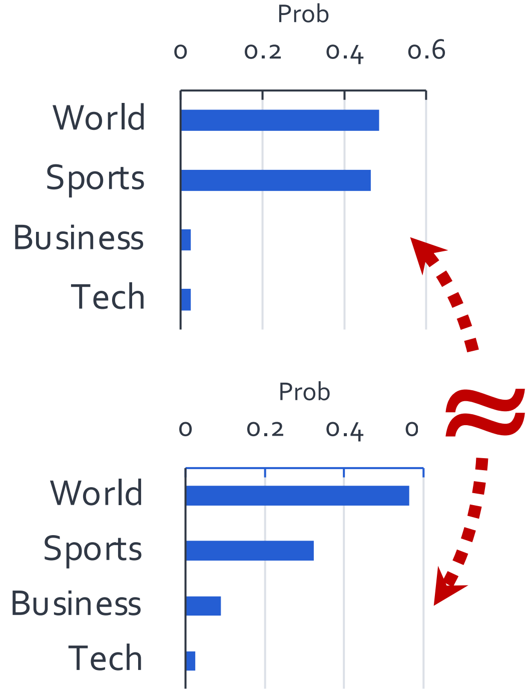
LM

```
Input: Rookie Taylor Wins Playoff at Tahoe
Output:
```

Gradient-descent

LM

Possible values:
`World, Sports, Business, Tech`

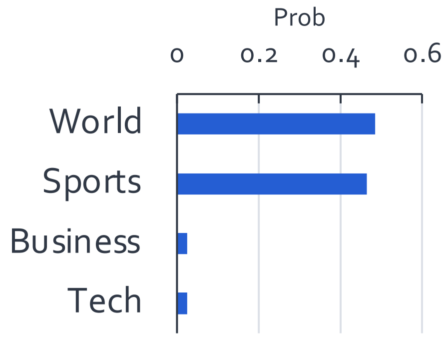


Can we take this
as an evidence for ICL ≈GD?

Full distributions over vocabulary are quite different!

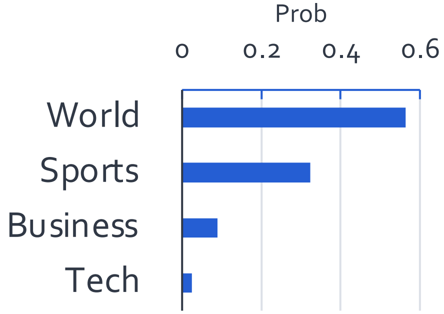
In-context

LM

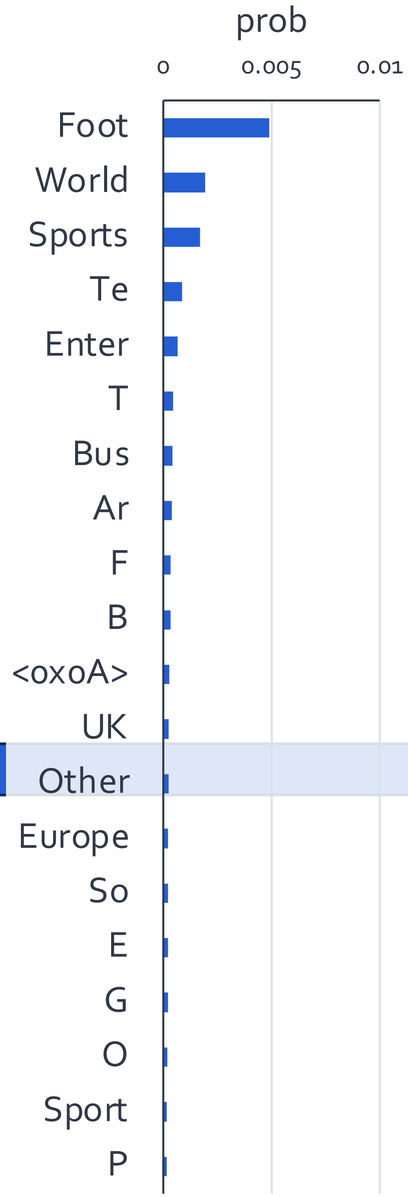


Gradient-descent

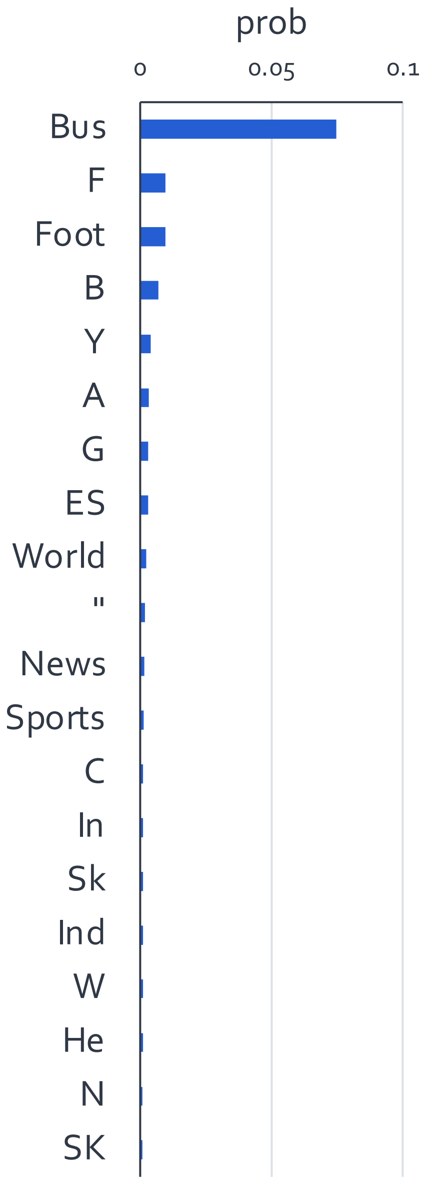
LM



In-context

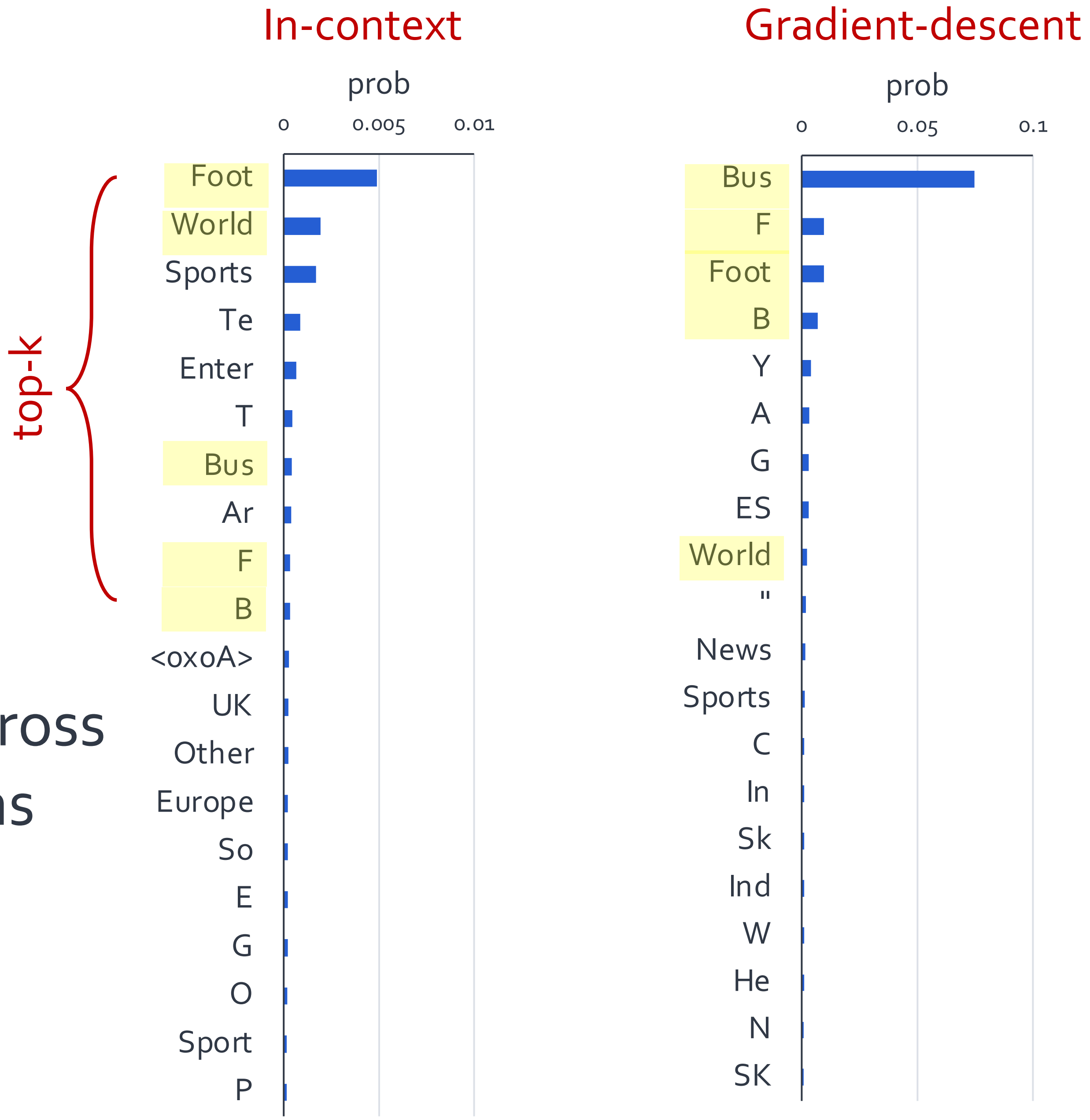


Gradient-descent



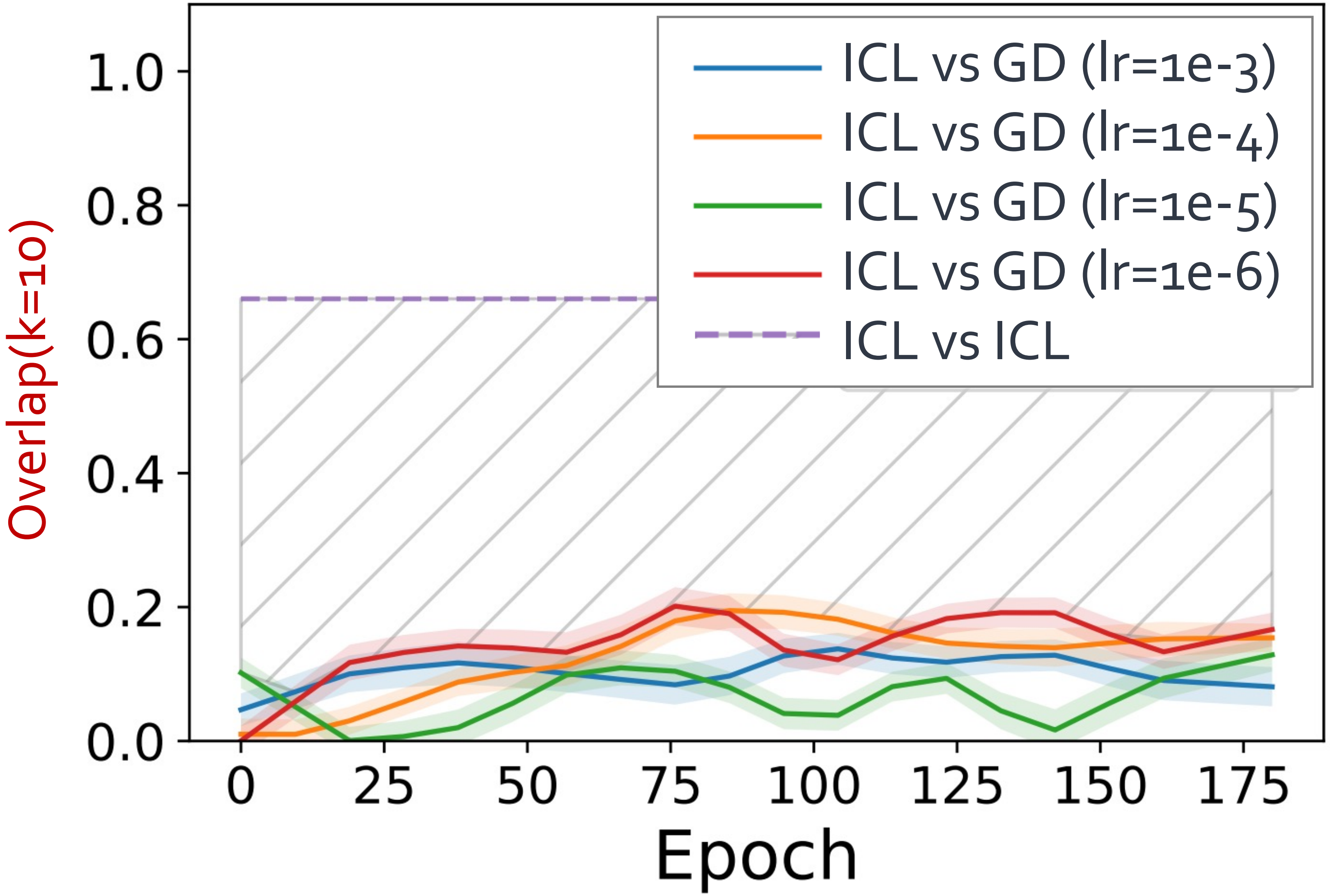
ICL vs GD modify LLM distributions differently.

Overlap(k=10)=50%



Overlap(k): % commonality across top-k tokens of two distributions

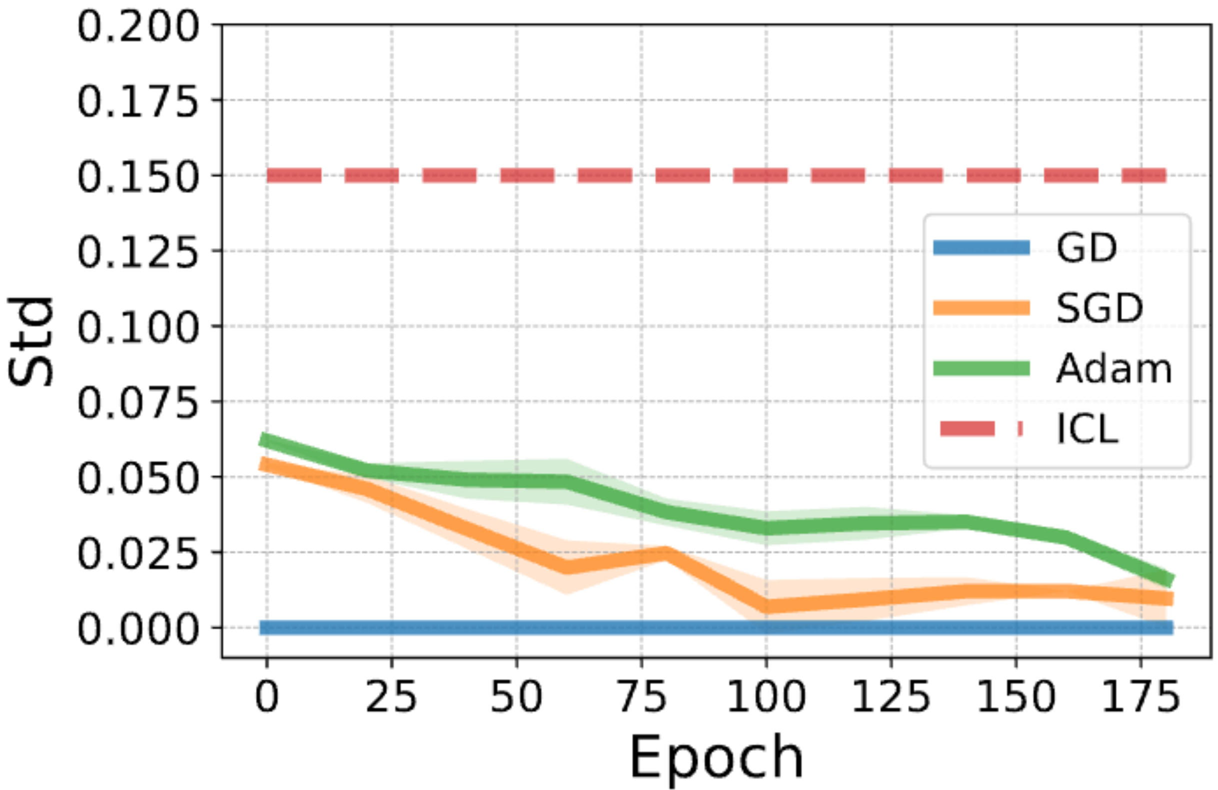
ICL vs GD modify LLM distributions differently.



ICL vs GD have different order-sensitivity.

- Prior research has demonstrated that ICL is highly sensitive to the order of in-context sample [Lu et al. 2022].
- GD and its variants is more order-stable (less STD).

Standard deviation token probabilities, for different choices of demonstrations.



Summary Thus Far

- The explanations of ICL based on GD are quite intriguing — do they hold in practice?

- In practice, we did not see any evidence that ICL simulates GD.
  - See the paper for more arguments and analysis.

- Note, we're not refuting it. It's left open for future research.
  - Deep inside, I believe that there must be a connection between ICL and optimization algorithms — we're just not looking at it right.

ICL remains understudied and elusive.

- ICL is the most important & mysterious phenomenon.
  - … we don't know how to explain it.
  - … and we are getting used to it.

- Many open problems:
  - Under what conditions does it emerge? (e.g., distributional properties)
  - Does ICL need natural language? Can it emerge, e.g., on brain signals?

ICL is likely what makes “alignment” effective.

- The success of LLMs in following instructions can be viewed from the lens of ICL.

- Being able to make LLMs adapt to various in-context demonstration was an early sign that LLMs can be controlled.

- To understand limits of controlling LLMs, we must understand limits of ICL.

Today

- Revisiting …

In-Context Learning

Alignment of chatbots

Today

- Revisiting …

In-Context Learning

Alignment of chatbots

Background

Language Modeling ≠ Following User Intents

Explain "space elevators" to a 6-year-old. → LM →

Explain gravity to a 6-year-old.
Explain black-holes to a 6-year-old.
Explain big bang to a 6-year-old.
....

LMs are not "aligned" with user intents [Ouyang et al., 2022].

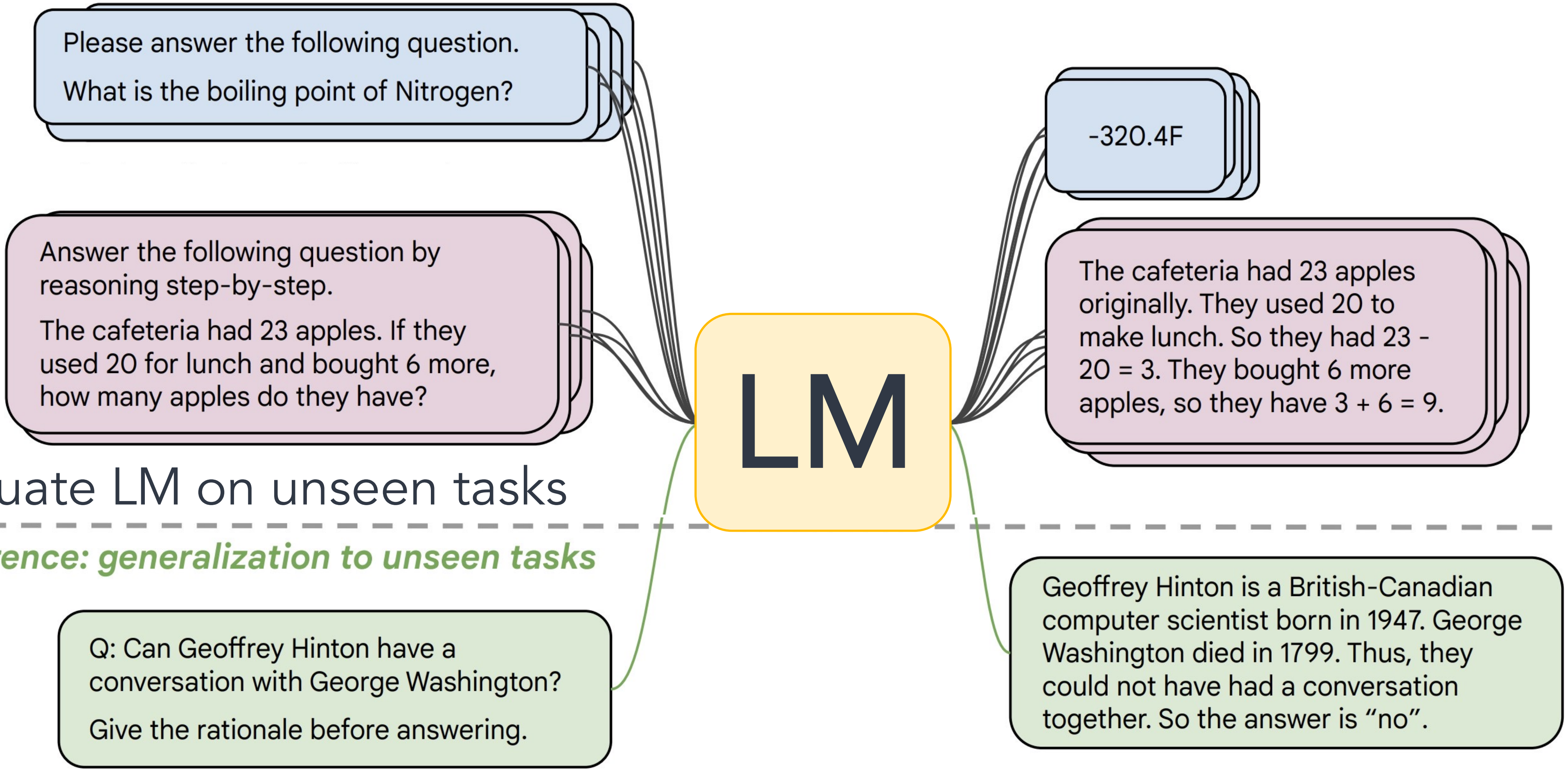
[Training language models to follow instructions with human feedback, Ouyang et al. 2022]

Background

How do we “align” LMs with our articulated intents?

Approach 1: Behavior Cloning (Supervised Learning)

1. Collect examples of (instruction, output) pairs across many tasks and finetune an LM



2. Evaluate LM on unseen tasks

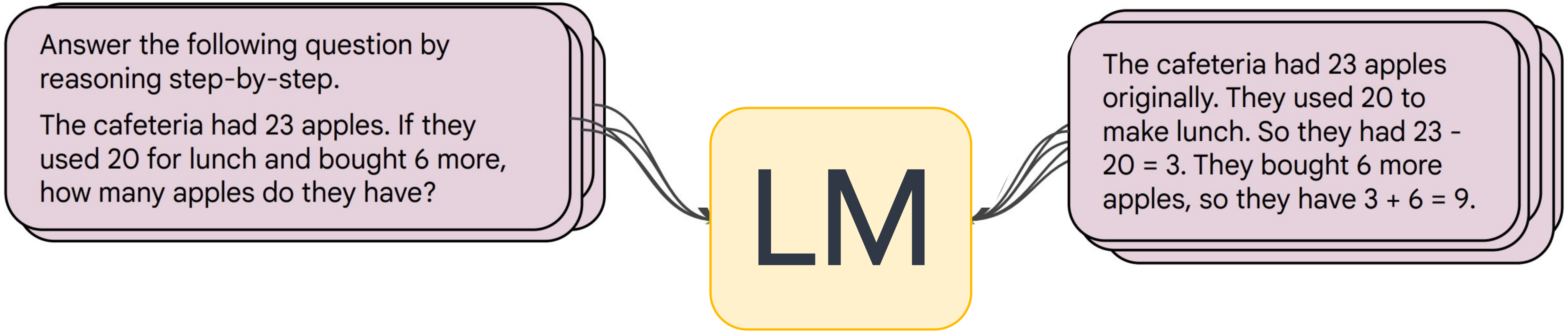
[McCann et al., 2019, Weller et al. 2020. Mishra et al. 2021; Wang et al. 2022, Sanh et al. 2022; Wei et al., 2022, Chung et al. 2022, many others]

Background

Approach 1: Behavior Cloning (Supervised Learning)

- Incentivizes word-by-word rote learning => limits creativity

- => The resulting models' generality/creativity is bounded by that of their supervision data.

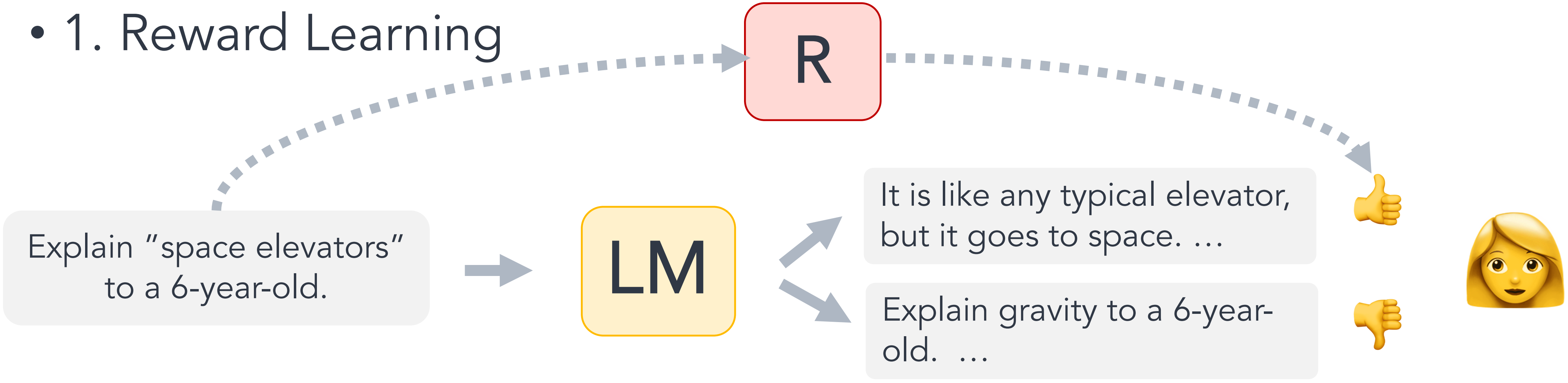


[McCann et al., 2019, Weller et al. 2020. Mishra et al. 2021; Wang et al. 2022, Sanh et al. 2022; Wei et al., 2022, Chung et al. 2022, many others]

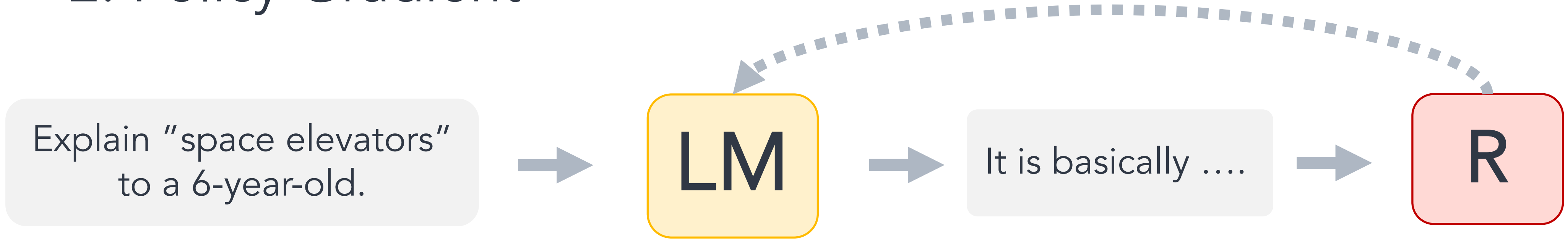
Background

Approach 2: RL w/ Ranking Feedback (RLHF)

- 1. Reward Learning



- 2. Policy Gradient



[Christiano et al. 2017; Stiennon et al. 2020; Ouyang et al., 2022]

Background

The overall recipe 👨‍🍳:

Pre-train → Align (instruct-tune) → Align (RLHF)

Background

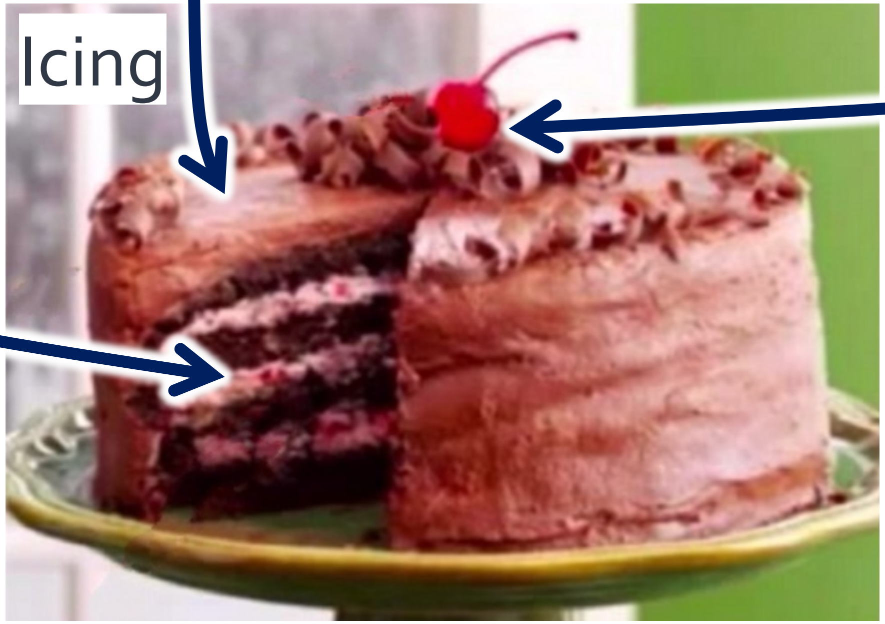
The overall recipe 🧑‍🍳:

Background

The overall recipe 👨‍🍳: Yann's Three-layered cake

Pre-train → Align (instruct-tune) → Align (RLHF)

Cake génoise



Cherry on the cake

Are these steps equally important?

Pre-train → Align (instruct-tune) → Align (RLHF)

Are these steps equally important?

Who should care?

1. **Product designers:** If you have $X million to build your best chatbot, how would you allocate it?
2. **Scientists:** Fundamentally, is this the ultimate pipeline?

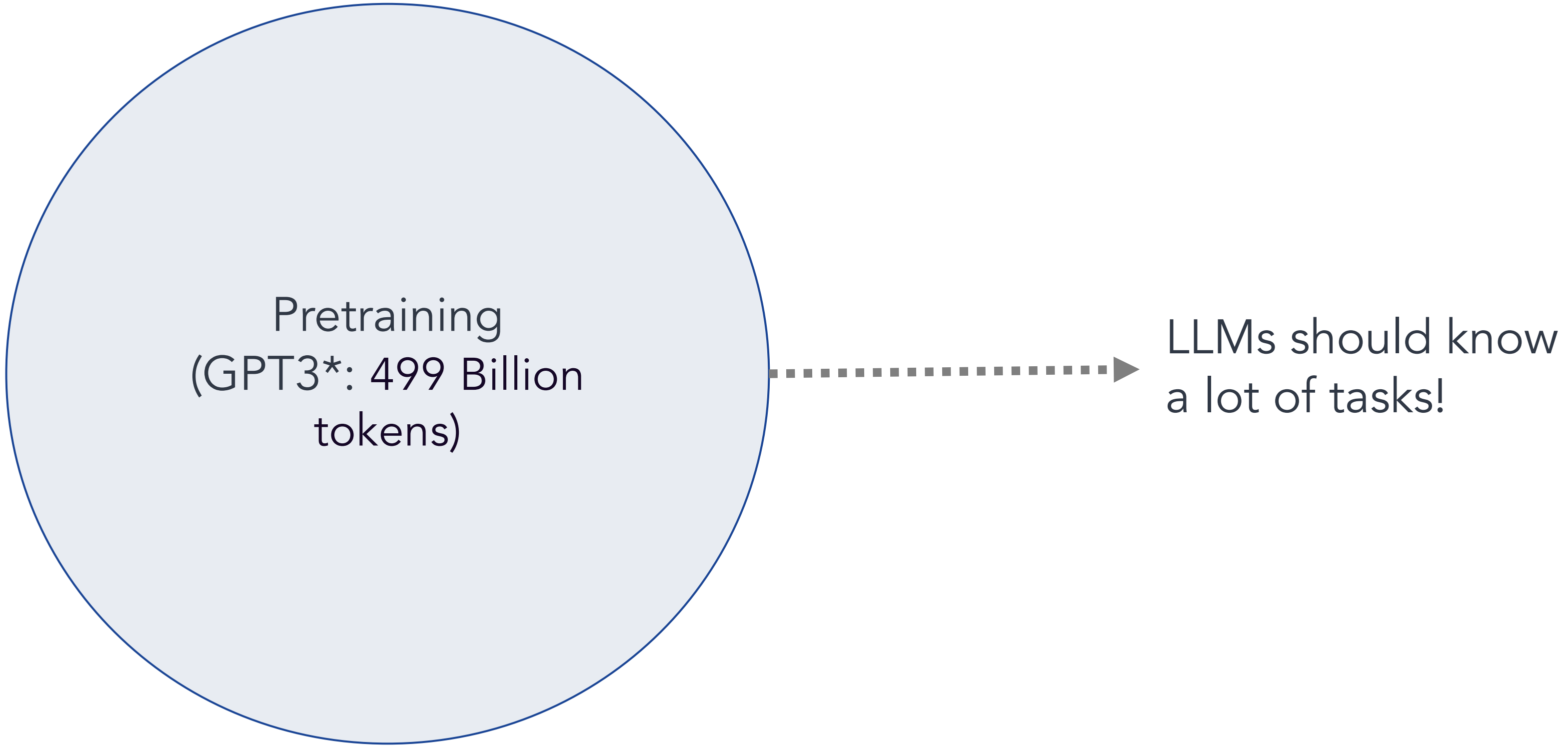
[Brown et al., 2020. GPT3, Ouyang et al., 2022. InstructGPT]

Are these steps equally important?

How far can we reduce the human annotations?

How far can we reduce the human annotations?

- **Idea:** we can bootstrap “instruction” from off-the-shelf LMs.
  - LMs have seen humans talk about their needs and goals.

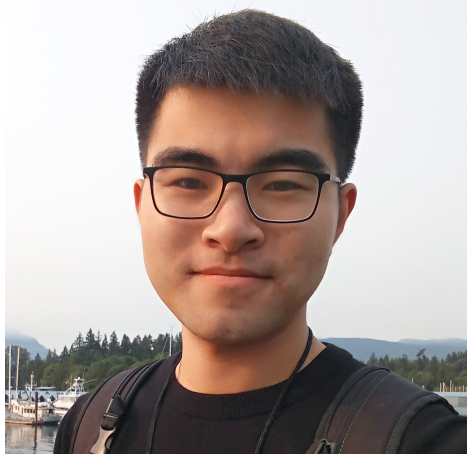


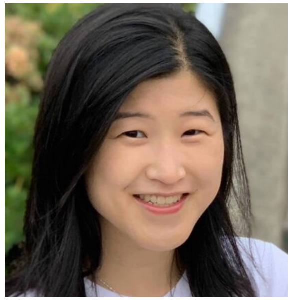
Warning: the paper is a year old!!

Self-Instruct:

Aligning Language Models w/ Self-Generated Instructions

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, Hannaneh Hajishirzi



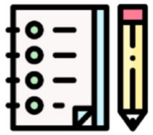




https://arxiv.org/abs/2212.10560

Get humans to write "seed" tasks ✍️

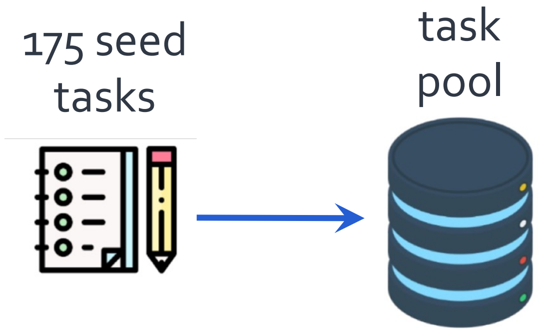
- I am planning a 7-day trip to Seattle. Can you make a detailed plan for me?
- Is there anything I can eat for breakfast that doesn't include eggs, yet includes protein and has roughly 700-1000 calories?
- Given a set of numbers find all possible subsets that sum to a given number.
- Give me a phrase that I can use to express I am very happy.

175 seed tasks



Put them your task bank 📦

- I am planning a 7-day trip to Seattle. Can you make a detailed plan for me?
- Is there anything I can eat for breakfast that doesn't include eggs, yet includes protein and has roughly 700-1000 calories?
- Given a set of numbers find all possible subsets that sum to a given number.
- Give me a phrase that I can use to express I am very happy.



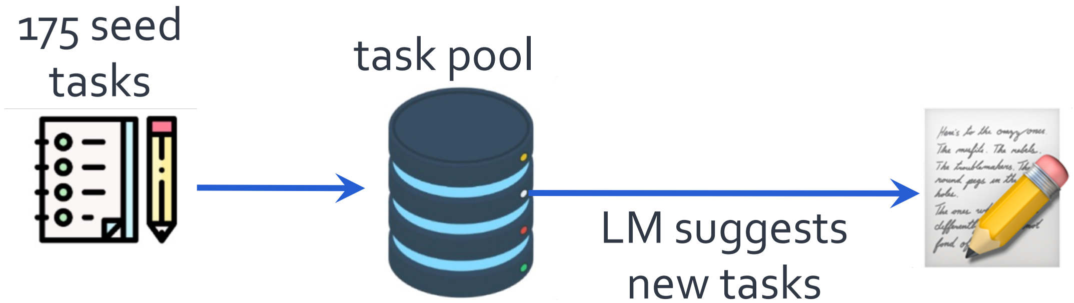
Sample and get LLM to expand it

- I am planning a 7-day trip to Seattle. Can you make a detailed plan for me?
- Is there anything I can eat for breakfast that doesn't include eggs, yet includes protein and has roughly 700-1000 calories?
- Given a set of numbers find all possible subsets that sum to a given number.
- Give me a phrase that I can use to express I am very happy.

LM

Pre-trained, but **not aligned yet**

- Create a list of 10 African countries and their capital city?
- Looking for a job, but it's difficult for me to find one. Can you help me?
- Write a Python program that tells if a given string contains anagrams.



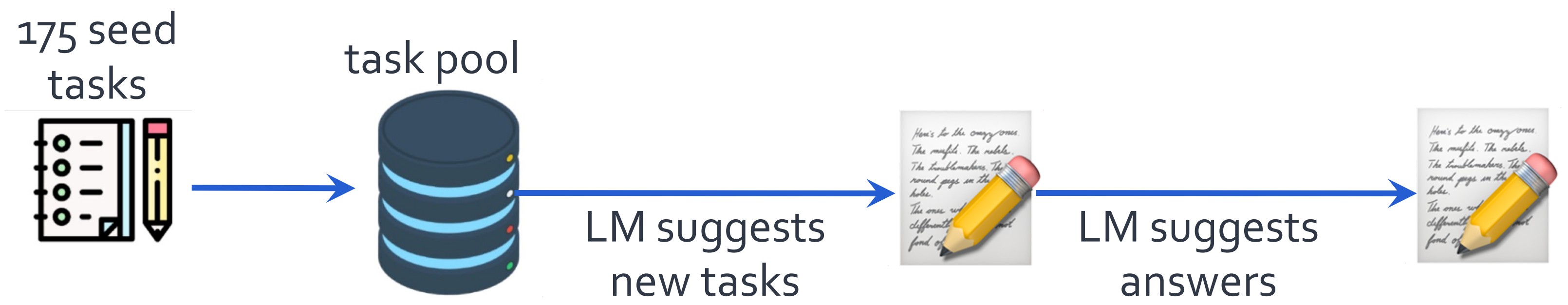
Get LLM to answers the new tasks

```
• Task: Convert the following temperature from Celsius to Fahrenheit.
• Input: 4 °C
• Output: 39.2 °F

• Task: Write a Python program that tells if a given string contains anagrams.
```

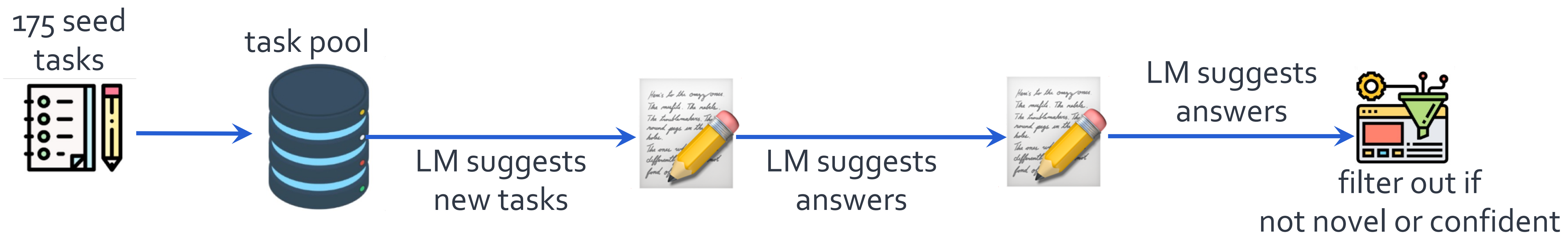
LM Pre-trained, but **not aligned yet**

```
• Input: -
• Output:
      def isAnagram(str1, str2): ...
```



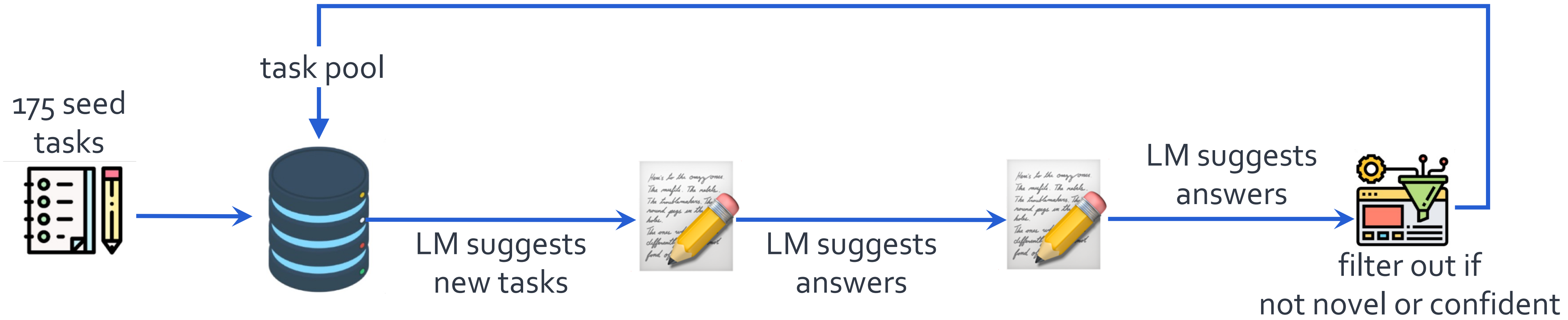
Filter tasks

- Drop tasks if LM assigns low probability to them.
- Drop tasks if they have a high overlap with one of the existing tasks in the task pool.
    - Otherwise, common tasks become more common — tyranny of majority.



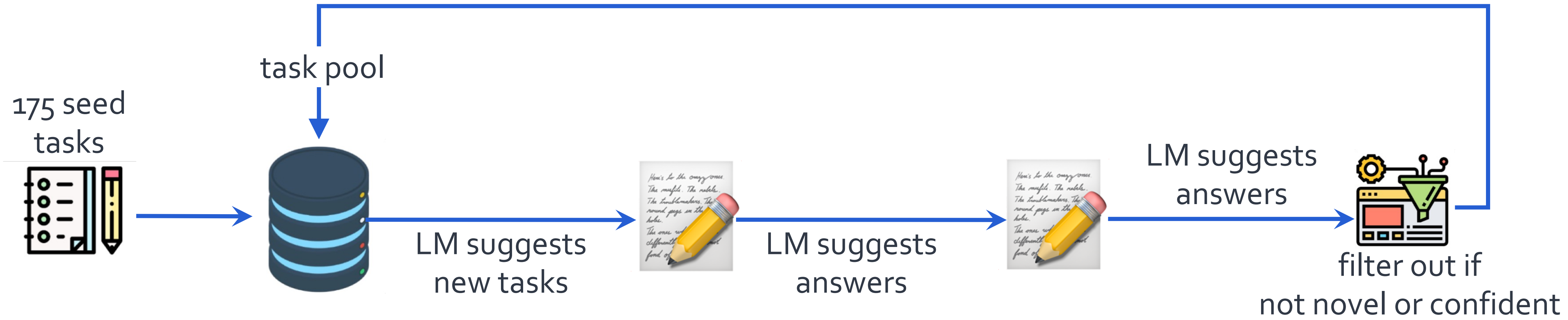
Close the loop

- Add the filtered tasks to the task pool.
- Iterate this process (generate, filter, add) until yield is near zero.

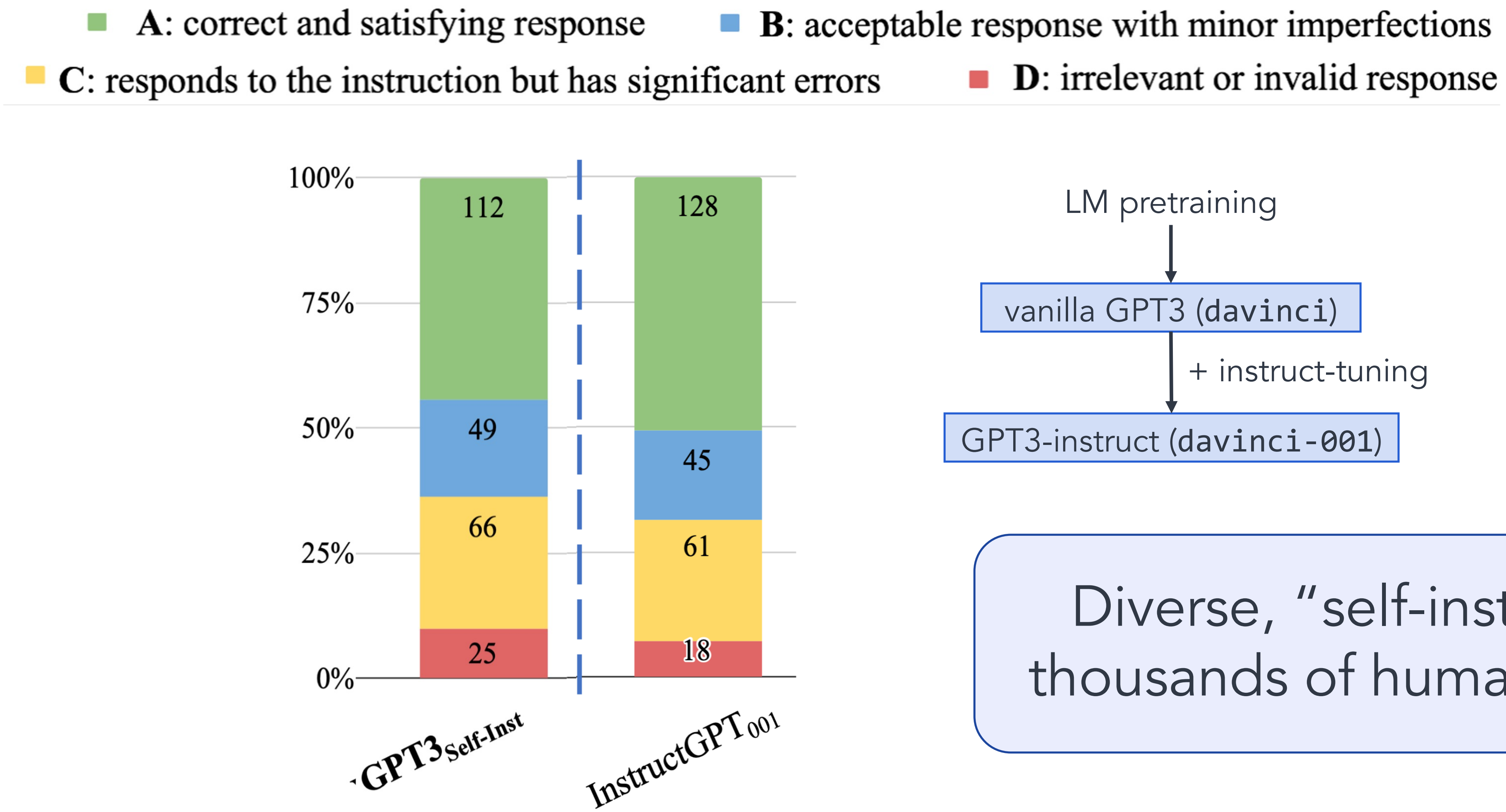


Self-Instructing GPT3 (base version)

- **Generate:**
  - GPT3 (“davinci” engine).
  - We generated 52K instructions and 82K instances.
  - API cost ~$600
- **Align:**
  - We finetuned GPT3 with this data via OpenAI API (2 epochs). **
  - API cost: ~$338 for finetuning



Evaluation on User-Oriented Instructions



LM pretraining

↓

vanilla GPT3 (`davinci`)

↓ + instruct-tuning

GPT3-instruct (`davinci-001`)

Diverse, "self-instruct" data ~ thousands of human-written data

Summary Thus Far

- There is a lot of room to reduce the reliance on human annotations in the “alignment” stage.
  - Well-read LLMs know a lot of our needs and demands.
  - Magic of “in-context learning” can surface these.

- Self-Instruct: Rely on creativity induced by LLMs themselves.
  - Lots of open-source adoption, but that’s not the point …

(* See also concurrent work: Unnatural-Instructions [Honovich et al. 2022] and Self-Chat [Xu et al. 2023])

The weight of "alignment" step

Fundamentally, what is the role of post hoc alignment (step #2/3)?

Step #1: Pre-train → Step #2/3: Align (RLHF or instruction-tune)

It's playing a small role — Lightly modify LM so it can articulate its existing knowledge of tasks.

(+ put guardrails for what it can articulate)

⟷

It's playing a big role — Teaching LM knowledge of new tasks.

Implications for how to invest

Fundamentally, what is the role of post hoc alignment (step #2/3)?

Step #1: Pre-train → Step #2/3: Align (RLHF or instruction-tune)

Make it more efficient, possibly with minimal human labor.

It's playing a small role — Lightly modify LM so it can articulate its existing knowledge of tasks.

(+ put guardrails for what it can articulate)

⟷

It ought to be annotation-intensive to teach the necessary knowledge.

It's playing a big role — Teaching LM knowledge of new tasks.

Implications for what comes out

Fundamentally, what is the role of post hoc alignment (step #2/3)?

Step #1: Pre-train → Step #2/3: Align (RLHF or instruction-tune)

Unexpected behaviors may "emerge".

It's playing a small role — Lightly modify LM so it can articulate its existing knowledge of tasks.

(+ put guardrails for what it can articulate)

↔

It will be as good as the alignment supervision.

It's playing a big role — Teaching LM knowledge of new tasks.

The weight of "alignment" step: My 2 cents

- Most of the heavy lifting is done via pre-training (unlabeled).

- Alignment to "instructions" (tuning/RLHF) is a light touch on LLMs.
  - Can (and should) be done more efficiently.

It's playing a small role — Lightly modify LM so it can articulate its existing knowledge of tasks.

(+ put guardrails for what it can articulate)

⟷

It's playing a big role — Teaching LM knowledge of new tasks.

RLHF is patchwork for lack of grounding

- RLHF teach LMs (ground) the communicative intent of users.
    - For example, what is intended by “summarize”? The act of producing a summary grounded in the human concept of "summary”.

- Not a panacea, but a short-term “band-aid” solution.

[Some remarks on Large Language Models, Goldberg 2023]

Alignment as a social process

- Can alignment emerge as a social experience?
- Internet also captures a subset of the world's interactive experiences.

The future is a cheesecake

- Future: A unifying process that encompasses various steps that are done separately today.

- The margins between alignment stages are getting murkier.
  - Using model itself for feedback and verification
  - Alignment during pre-training (Korbak et al. 2023)
  - Building bridges between supervised learning and RL (see DPO vs. RLHF)
  - …

Pre-train → Align (instruct-tune) → Align (RLHF)

The future is a cheesecake

- Future: A unifying process that encompasses various steps that are done separately today.

- Yann's framework was good for getting a system off the ground.

- Now that we are moving to interactive setups, alignment and pre-training will be a continual process. Systems that :
  - Adaptively change to our needs and habits;
  - Seamlessly pick up on implicit reward;
  - ….

Thanks!