

CSE 156 | Lecture 10: Prompting; Instruction Tuning

Ndapa Nakashole

October 31, 2024

Administrative matters

- ▶ **PA2:** due in a week
- ▶ **Quiz 1:** was due yesterday

Today

- ① Prompting
- ② Instruction Tuning
- ③ Alignment of LLMs

Recap: GPT-2 decoder only; new datatet, towards zero-shot transfer learning

Tasks are present in the data

"I'm not the cleverest man in the world, but like they say in French: **Je ne suis pas un imbecile [I'm not a fool].**

In a now-deleted post from Aug. 16, Soheil Eid, Tory candidate in the riding of Joliette, wrote in French: "**Mentez mentez, il en restera toujours quelque chose,**" which translates as, "**Lie lie and something will always remain.**"

"I hate the word '**perfume**','" Burr says. 'It's somewhat better in French: '**parfum**'.'

If listened carefully at 29:55, a conversation can be heard between two guys in French: "**-Comment on fait pour aller de l'autre côté? -Quel autre côté?**", which means "**- How do you get to the other side? - What side?**".

If this sounds like a bit of a stretch, consider this question in French: **As-tu aller au cinéma?**, or **Did you go to the movies?**, which literally translates as Have-you to go to movies/theater?

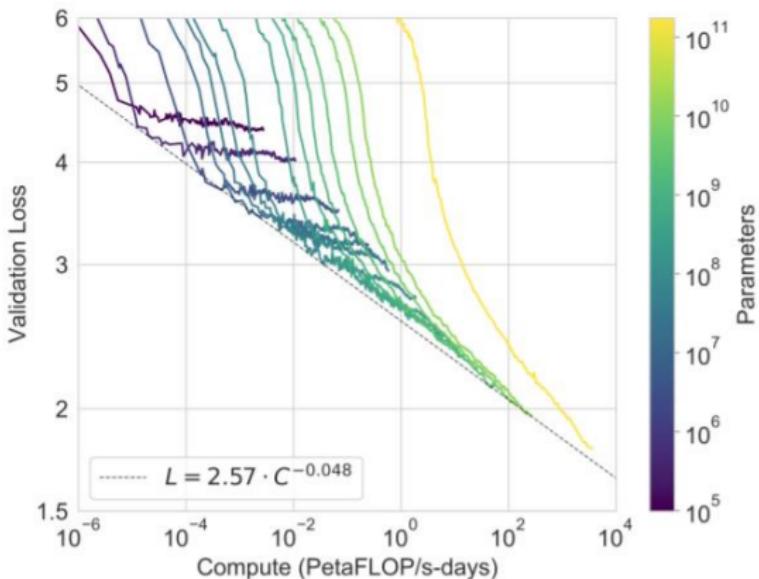
Recap: Decoding Strategies

- ▶ LMs place a distribution $P(x_i | x_1, \dots, x_{i-1})$. How do we generate text from these?
- ▶ Maximization-based decoding:
 - **Greedy decoding**
 - **Beam search**
- ▶ Sampling-based decoding:
 - **Random sampling**
 - **Nucleus sampling**
 - **Top- k sampling** (optionally with temperature)

Recap: Scaling Laws

Simple, predictive "laws" for behaviors of LMs

- ▶ **Old and unpleasant:** tune hyperparameters on big models
- ▶ **New and exciting:** tune on small models, extrapolate to large ones



A question

Do autoregressive LMs simply predict the next token?

- ➊ Yes, that is all they do
- ➋ Well, they predict scores over the entire vocabulary at each step. We then use those scores to predict the next token
- ➌ And, actually, they also represent data in their internal and output representations
- ➍ But saying they simply predict the next token might be best in terms of communication with the public

Prompting, In-Context Learning

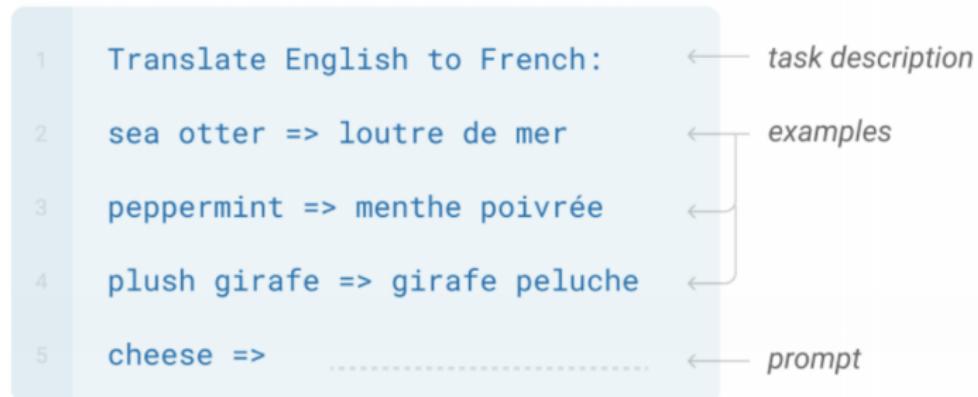
Pre-GPT-3: Fine-tuning

- ▶ **Fine-tuning:** train on a small dataset for a specific task; normal way for BERT,
...
▶ **Problem:** requires applying a parameter update on every example
▶ **This is super expensive with 175B parameters**



GPT-3: Few-shot Learning

- ▶ GPT-3 proposes an alternative: **in-context learning**; Just uses the off-the-shelf model no gradient updates



Brown et al. (2020)

- ▶ This procedure depends heavily on the examples you pick as well as the prompt ("Translate English to French")

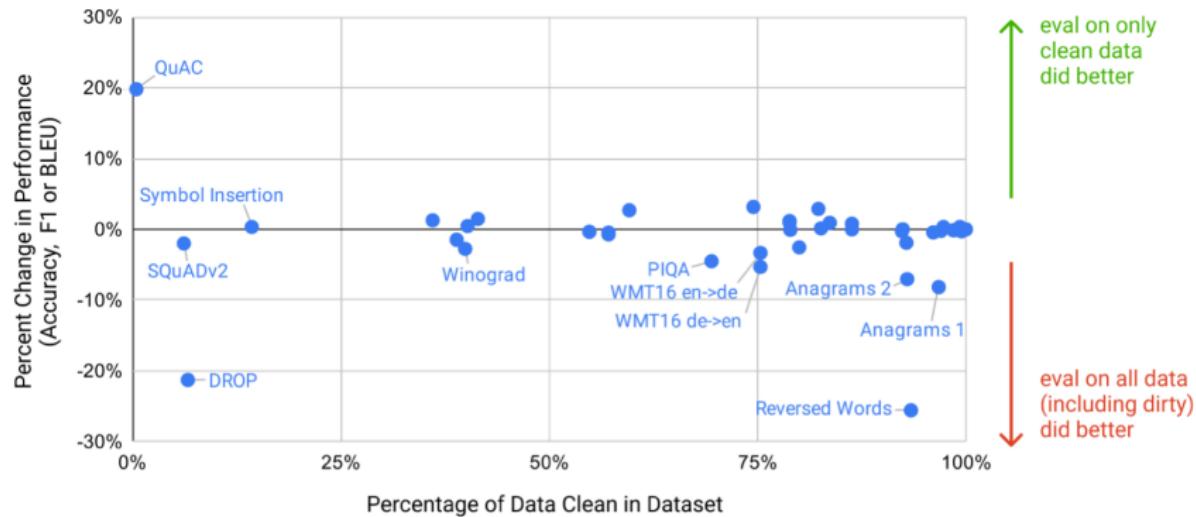
GPT-3

	SuperGLUE Average	BoolQ Accuracy	CB Accuracy	CB F1
Fine-tuned SOTA	89.0	91.0	96.9	93.9
Fine-tuned BERT-Large	69.0	77.4	83.6	75.7
GPT-3 Few-Shot	71.8	76.4	75.6	52.0
	WiC Accuracy	WSC Accuracy	MultiRC Accuracy	MultiRC F1
Fine-tuned SOTA	76.1	93.8	62.3	88.2
Fine-tuned BERT-Large	69.6	64.6	24.1	70.0
GPT-3 Few-Shot	49.4	80.1	30.5	75.4

- ▶ Sometimes impressive, sometimes bad
- ▶ Results on other datasets are equally mixed - but still strong for a few-shot model!

Data contamination

Analysis: did the model just memorize everything?



- ▶ Overlap can be large (e.g., many reading comprehension articles come from wikipedia)
- ▶ Results on clean part of the benchmark don't change much

Prompting to perform Tasks

Zero-shot Prompting for Text Classification

- ▶ Single unlabeled datapoint x , want to predict label y
 $x = \text{The movie's acting could've been better}$
- ▶ Wrap x in a template we call a **verbalizer** v

Prompt:

Review: The movie's acting could've been better
Out of positive, negative, or neutral, this review is ____



negative

Zero-shot Prompting for Text Classification

- ▶ Single unlabeled datapoint x , want to predict label y
 $x = \text{The movie's acting could've been better}$
- ▶ Wrap x in a template we call a **verbalizer** v

Prompt:

Review: The movie's acting could've been better

On a 1 to 4 star scale, the reviewer would probably give
this movie ___



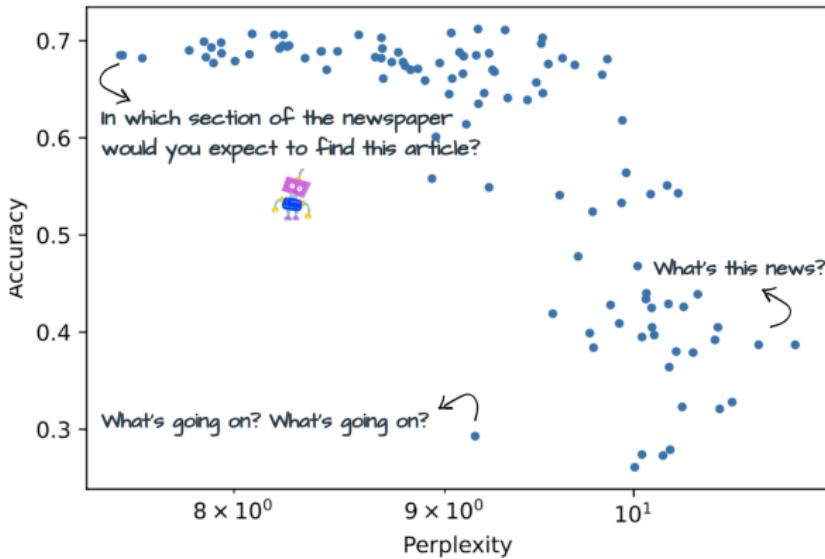
2 stars

Ways to do classification

- ▶ Generate from the model and read off the generation
 - What if you ask for a star rating and but LM is too **verbose**
"I would rate this movie 2 stars"?
- ▶ Compare probabilities: "Out of positive, negative, or neutral,
this review is ____"
 - Compare: $P(\text{positive} \mid \text{context})$, $P(\text{neutral} \mid \text{context})$,
 $P(\text{negative} \mid \text{context})$
 - This constrains the model to only output a valid answer; can
normalize probabilities to get a distribution

Variability in Prompts

- ▶ Plot: large number of prompts produced by manual writing, paraphrasing, backtranslation, etc.
- ▶ x -axis: perplexity of the prompt. How natural is it? How much does it appear in the pre-training data?
- ▶ y -axis: task performance



Variability in Prompts

- ▶ OPT-175B: average of best 50% of prompts is much better than average over all prompts

Task	Avg Acc	Acc 50%
Antonyms	—	—
GLUE Cola	47.7	57.1
Newspop	66.4	72.9
AG News	57.5	68.7
IMDB	86.2	91.0
DBpedia	46.7	55.2
Emotion	16.4	23.0
Tweet Offensive	51.3	55.8

Few-shot prompting

- ▶ Provide a few examples of the task together with the instruction

Prompt:

Please classify movie reviews as 'positive' or 'neutral' or 'negative'.

Input: I really don't like this movie.

Output: negative

Input: This movie is great!

Output: positive

Input: The movie's acting could've been better.

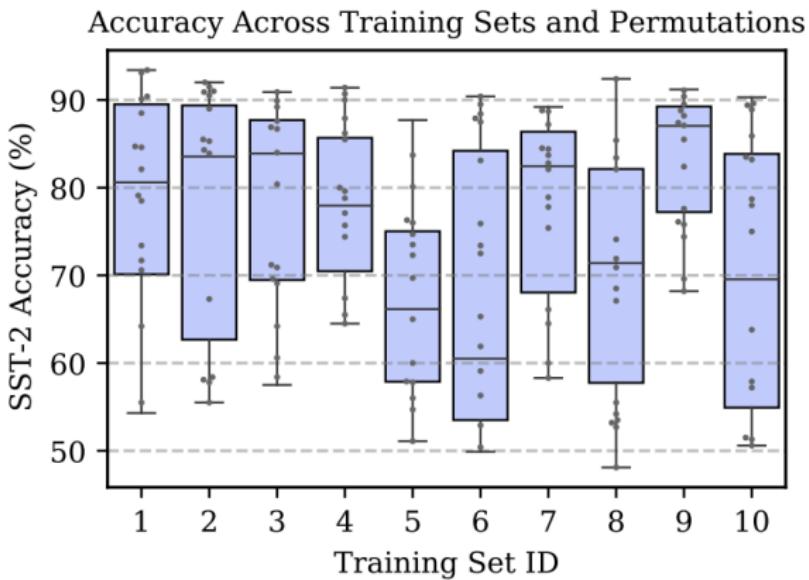
Output:



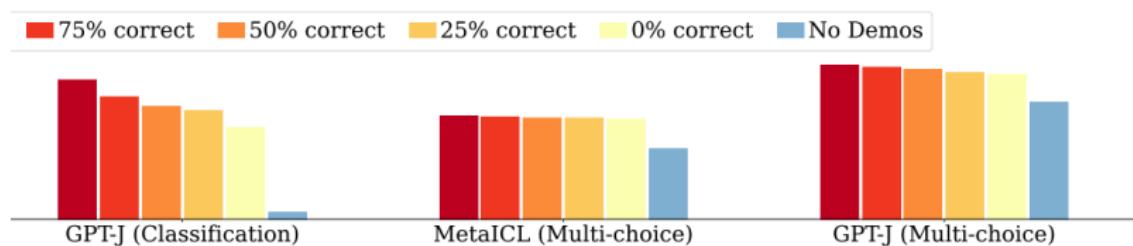
negative

Few-shot Prompting: What can go wrong?

- ▶ x -axis: different collections of train examples.
- ▶ y -axis: sentiment accuracy. Boxes represent results over different permutations of the data
- ▶ Performance: varies even across permutations of training examples



Rethinking Demonstrations: Effects are Sometimes Counter-intuitive



- ▶ Surprising result: how necessary are the demonstrations?
- ▶ Using random labels does not substantially decrease performance??
- ▶ Mislabeled demonstrations better than having no demonstrations
 - Suggests that the form of the demonstrations is partially responsible for in-context learning

Prompting: Summary

- ▶ Zero- and few-shot prompting are very powerful ways of specifying new tasks at inference time
- ▶ For zero-shot: prompt can make a huge difference
- ▶ For few-shot: number and order of the examples matter, prompt matters a bit less

Chain-of-thought Prompting

Chain of Thought (CoT) Prompting

- ▶ Give the model time to "think" by getting it to explain its reasoning before generating an answer

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. X

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: 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 $3 + 6 = 9$. The answer is 9. ✓

[Wei et al., 2022; also see Nye et al., 2021]

Zero-Shot Chain of Thought Prompting

- ▶ Just add a prompt that encourages the model to explain decisions can induce reasoning

(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The answer is 8. ✗

(b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are $16 / 2 = 8$ golf balls. Half of the golf balls are blue. So there are $8 / 2 = 4$ blue golf balls. The answer is 4. ✓

(c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: The answer (arabic numerals) is

(Output) 8 ✗

(d) Zero-shot-CoT (Ours)

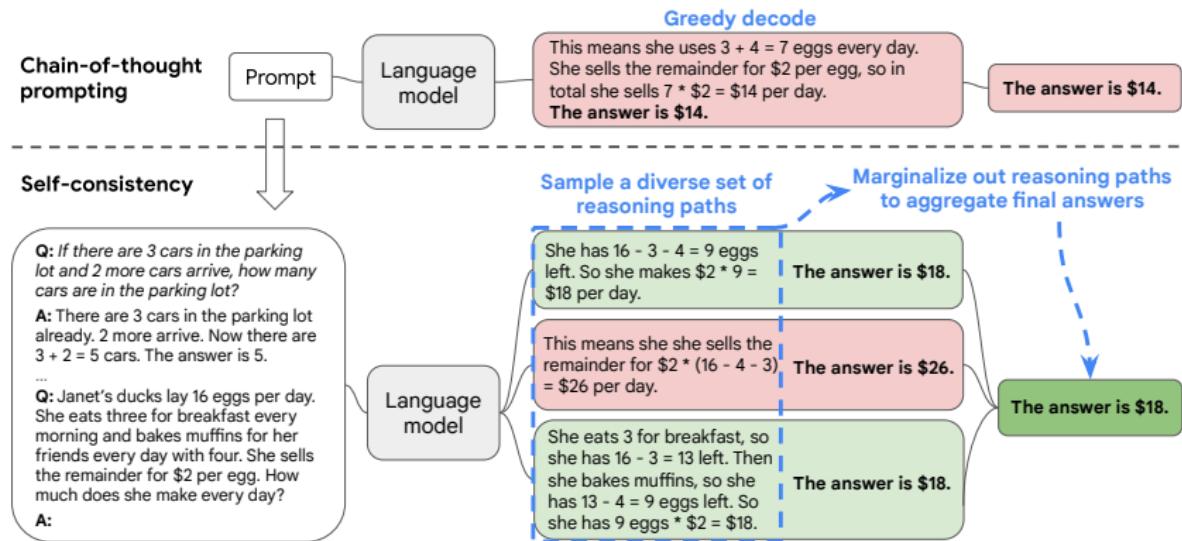
Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: Let's think step by step.

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

Note: GPT models reason even w/o specific instructions now (probably due to instruction tuning)

Self-Consistency in Chain of Thought Prompting



- ▶ Ensembling across multiple outputs (either zero-shot or few-shot)
- ▶ Gains on several other math datasets and text tasks

Chain of Thought Prompting: Summary

- ▶ Chain of thought prompting can be a powerful way to get models to reason
 - E.g., mathematical reasoning and textual question answering with multiple steps
- ▶ Self-consistency can improve performance

Prompt-Engineering Paradigms

- ▶ Manual
 - Configure a manual template based on the characteristics of the task
- ▶ Automated Prompt Engineering
 - Paraphrasing, back translation
 - Prompt tuning in the embedding space

Manual Prompt Engineering

- ▶ Instructions should be clear, concise and easy to understand
- ▶ Good examples:

<https://www.promptingguide.ai/introduction/tips>

less precise

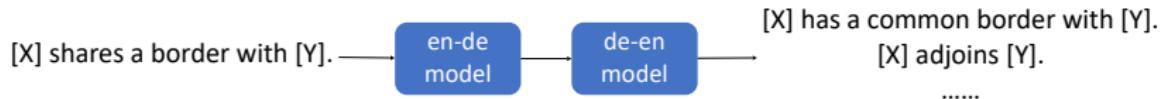
Explain the concept prompt engineering. Keep the explanation short, only a few sentences, and don't be too descriptive.

more precise

Use 2-3 sentences to explain the concept of prompt engineering to a high school student.

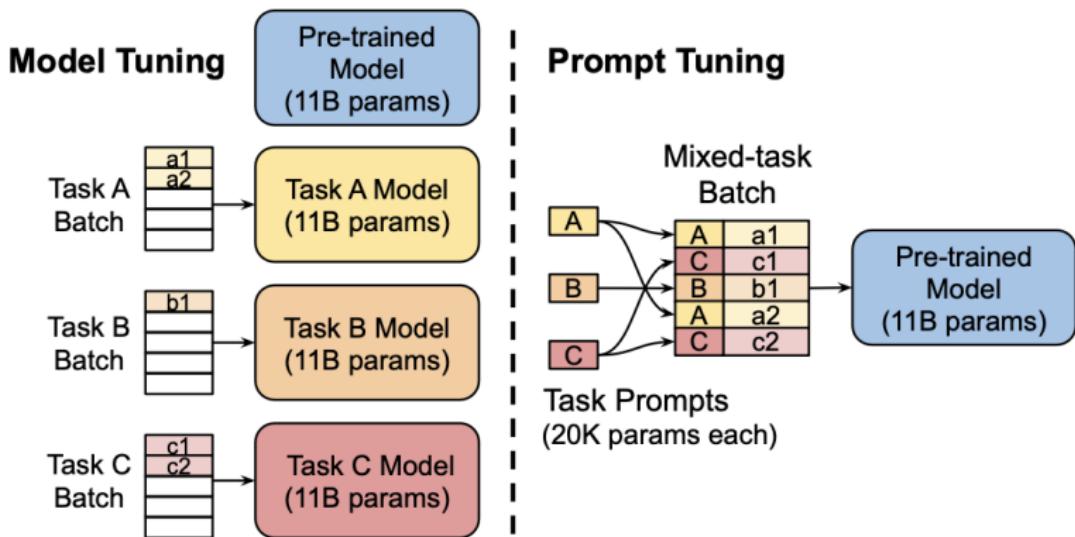
Prompt Paraphrasing

- ▶ Paraphrase an existing prompt to get other candidates (Jiang et al. 2019)
- ▶ Use back translation



Prompt Tuning in Embedding Space

- ▶ Optimize the embeddings of a prompt, instead of the words
- ▶ Only requires storing a small task-specific prompt for each task



Instruction Tuning: From Language Models to AI Assistants

```
In [5]: prompt = """  
Tell me about student life at the University of California, San Diego  
which is located in Dallas, Texas.  
"""  
  
response = get_completion(prompt)  
print(response)
```

I'm sorry, but there is an error in your question. The University of California, San Diego is actually located in La Jolla, California, not Dallas, Texas.

As an AI language model, I don't have personal experience of student life at UC San Diego, but I can provide some general information. UC San Diego is a large public research university with a diverse student body of over 38,000 undergraduate and graduate students. The campus is known for its strong academic programs in science, engineering, and social sciences, as well as its beautiful coastal location.

There are over 600 student organizations on campus, ranging from cultural and ethnic groups to academic and professional organizations, sports clubs, and more. Students can also participate in intramural sports, fitness classes, and outdoor adventures through the uni

Base LLM

Predicts next word, based on
text training data

Once upon a time, there was a unicorn

Base LLM

Predicts next word, based on
text training data

Once upon a time, there was a unicorn
that lived in a magical forest with
all her unicorn friends

Base LLM

Predicts next word, based on
text training data

Once upon a time, there was a unicorn
that lived in a magical forest with
all her unicorn friends

What is the capital of France?

Language modeling \neq assisting users

Base LLM

Predicts next word, based on
text training data

Once upon a time, there was a unicorn
that lived in a magical forest with
all her unicorn friends

What is the capital of France?

What is France's largest city?

What is France's population?

What is the currency of France?

Problem: The base LLM is not trained to follow instructions, and may not always generate the desired output.

Instruction Tuned LLM

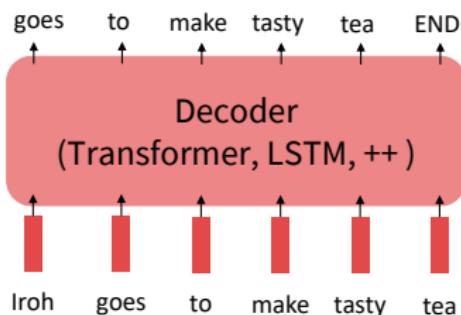
What is the capital of France?

The capital of France is Paris.

Recall Pretrain/Finetune paradigm ...

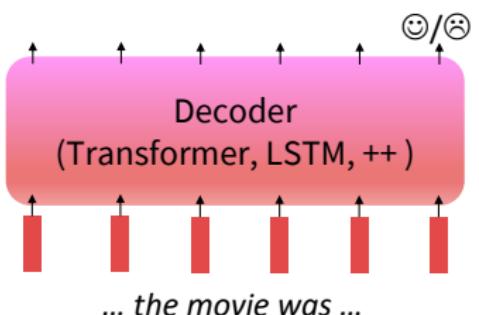
Step 1: Pretrain (on language modeling)

Lots of text; learn general things!



Step 2: Finetune (on your task)

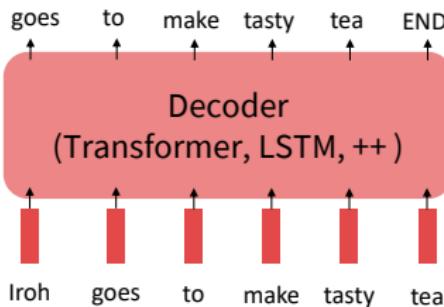
Not many labels; adapt to the task!



Scaling up Finetuning

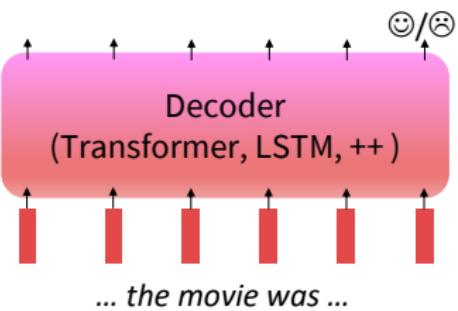
Step 1: Pretrain (on language modeling)

Lots of text; learn general things!



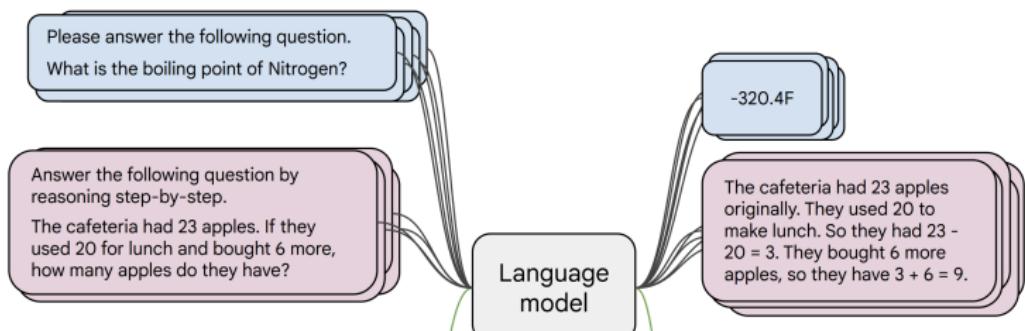
Step 2: Finetune (on many tasks)

Not many labels; adapt to the tasks!



Instruction finetuning

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



Evaluate on **unseen tasks**

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

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".

[FLAN-T5; [Chung et al., 2022](#)]

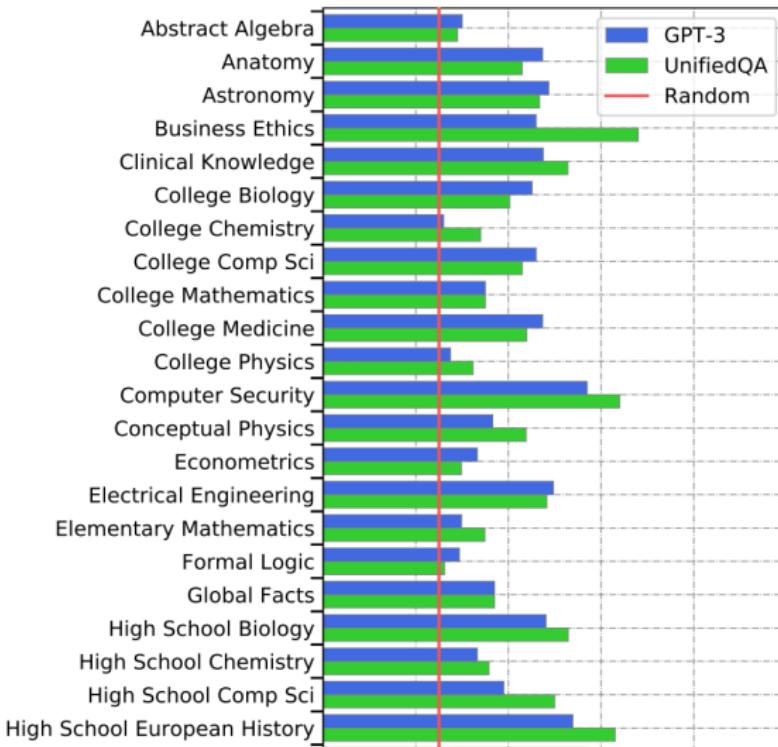
Instruction finetuning pretraining?

- ▶ As always, **data + model scale** is key for this to work!
- ▶ E.g., the SuperNaturalInstructions dataset contains over **1.6 K tasks, 3M+ examples**



Aside: new benchmarks for multitask LMs

- ▶ Massive Multitask Language Understanding (MMLU) [Hendrycks et al., 2021]
- ▶ New benchmarks for measuring LM performance on 57 diverse knowledge intensive tasks



Examples from MMLU

Astronomy

What is true for a type-Ia supernova?

- A. This type occurs in binary systems.
- B. This type occurs in young galaxies.
- C. This type produces gamma-ray bursts.
- D. This type produces high amounts of X-rays.

Answer: A

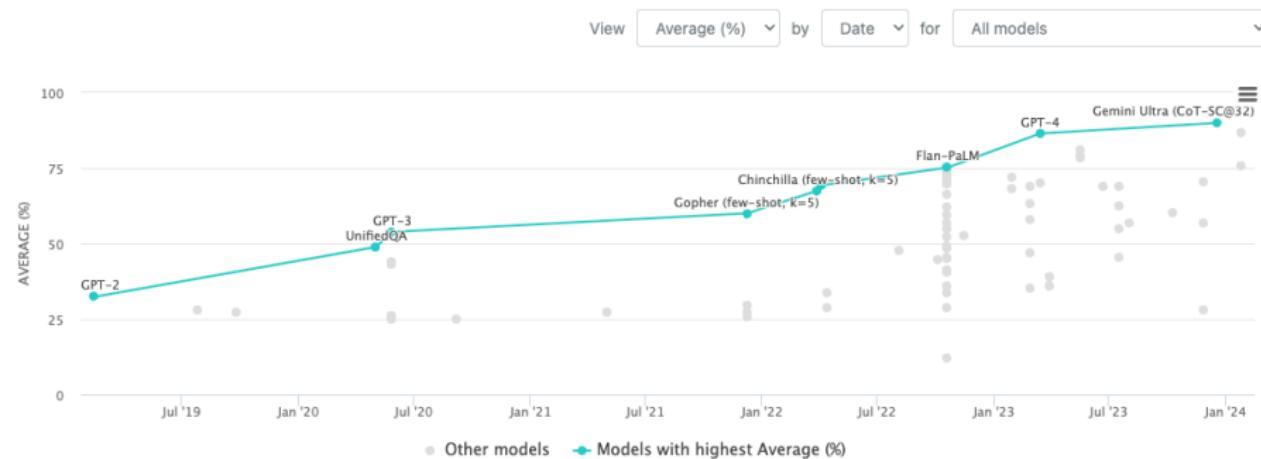
High School Biology

In a population of giraffes, an environmental change occurs that favors individuals that are tallest. As a result, more of the taller individuals are able to obtain nutrients and survive to pass along their genetic information. This is an example of

- A. directional selection.
- B. stabilizing selection.
- C. sexual selection.
- D. disruptive selection

Answer: A

Progress on MMLU



Rapid, impressive progress on challenging knowledge-intensive benchmarks

Aside: new benchmarks for multitask LMs

BIG-Bench [[Srivastava et al., 2022](#)]

200+ tasks, spanning:



https://github.com/google/BIG-bench/blob/main/bigbench/benchmark_tasks/README.md

Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models

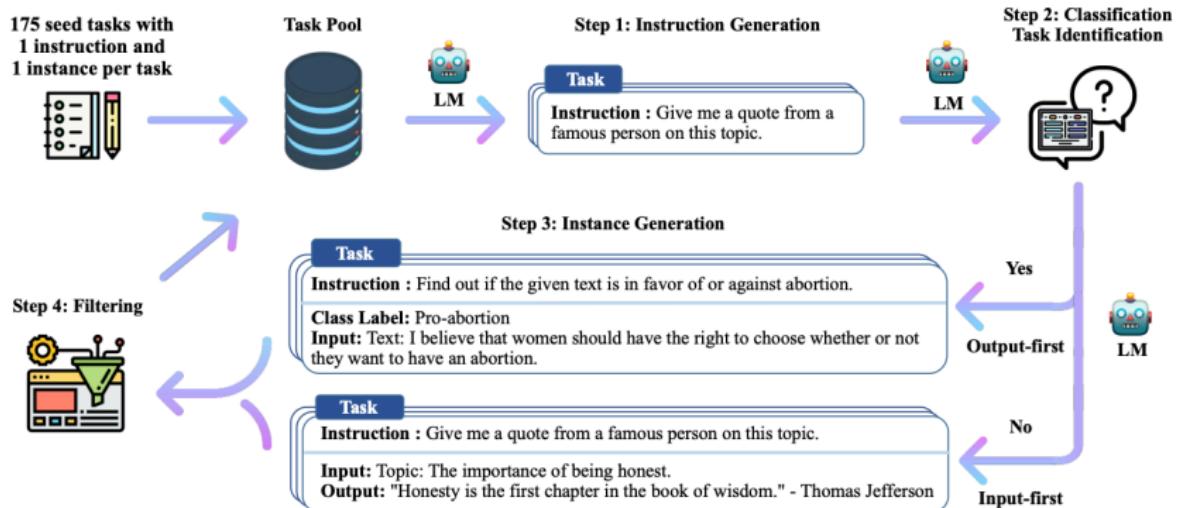
alphabetical author list:^{*}

Instruction Tuning: ongoing directions & open questions

- ▶ Expanding task diversity more aggressively through **synthetic data generation**
- ▶ **Open question:** What level of progress can be achieved without costly human response demonstrations or ratings?
 - Crucial **narrow the gap** between **public and non-public research** in this area

Self-Instruct/Alpaca

Automatically generate instruction tuning datasets (Wang et al. 2023); used to instruct-tune Llama into Alpaca



- Fine-tune Llama on 52k outputs with answers generated by text-davinci-003

Yizhong Wang et al. (2023) Self-Instruct

Ronen Taori et al. (2023) Alpaca

Examples of Open Instruction Tuned models

FLAN-T5: [huggingface/google/flan-t5-xxl](https://huggingface.co/google/flan-t5-xxl)

- ▶ **Architecture:** Encoder-decoder model, based on T5
- ▶ **Parameter Count:** 11B parameters
- ▶ Flan-T5 [Chung et al., 2020]: T5 models finetuned on 1.8K additional tasks

LLaMa Chat models, e.g.:

[huggingface/meta-llama/Llama-2-70b-chat-hf](https://huggingface.co/meta-llama/Llama-2-70b-chat-hf)

- ▶ **Architecture:** Decoder-only model
- ▶ **Parameter Count:** 70B parameters

Limitations of Instruction Tuning: RLHF

- ▶ **Data Collection Cost:** Collecting high-quality, ground-truth data for instruction-tuned tasks is costly.
- ▶ **Ambiguity in Open-Ended Tasks:** Many tasks, such as creative or open-ended generation, lack a single "correct" answer.
Example: "*Write me a story about a dog and her pet grasshopper*"
- ▶ **Uniform Error Penalization:** In standard language modeling, all token-level errors are penalized equally.
However, *some mistakes impact the quality and coherence of outputs more than others.*

Optimizing for Human Preferences

► RLHF

- Explicitly attempt to satisfy human preferences using RL

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

...
overturn 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
prone to
earthquakes and
wildfires.

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

Want to maximize the expected reward of samples from our LM:

Reinforcement Learning (RL) for Instruction Following

- ▶ In Reinforcement Learning (RL), models receive instructions without human-provided answers.
- ▶ Models generate their own answers based on the given instructions.
- ▶ Answers are then evaluated by a scoring mechanism (e.g., human feedback) which provides feedback on answer quality.
 - The primary goal is for the model to learn how to produce answers that achieve high scores.

InstructGPT: scaling up RLHF

1) Instruction Finetuning, 2 and 3): maximize reward

Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

Explain the moon landing to a 6 year old

A labeler demonstrates the desired output behavior.

Some people went to the moon...

This data is used to fine-tune GPT-3 with supervised learning.

SFT

📄📄📄

Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

Explain the moon landing to a 6 year old

A Explain gravity...
B Explain mor...

C Moon is natural satellite of...
D People went to the moon...

A labeler ranks the outputs from best to worst.

D > C > A = B

This data is used to train our reward model.

RM

D > C > A = B

Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

Write a story about frogs

The policy generates an output.

PPO

Once upon a time...

The reward model calculates a reward for the output.

RM

The reward is used to update the policy using PPO.

r_k

Why both instruction-tuning & RLHF?

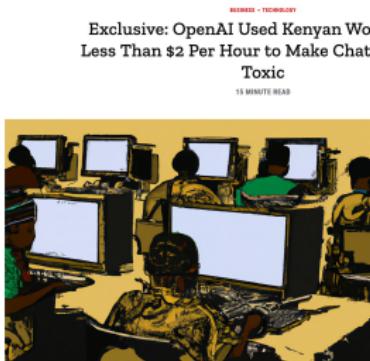
Instruction-tuning:

- ▶ Only **positive feedback** is provided through correct answers to questions
- ▶ Encourages **replication of provided answers**, which may lead to untruthful “guesses.”
- ▶ When the model lacks knowledge, it may either **memorize or fabricate** responses.
- ▶ Can inadvertently teach the model to “**lie**” by making up answers.

RLHF:

- ▶ Allows for both **positive and negative feedback** (more comprehensive learning)
- ▶ Model can **generate answers** and receive feedback on their correctness
- ▶ Reduces tendency to **fabricate answers**
- ▶ Helps models learn to **rely on internal knowledge** or **appropriately abstain from answering** when uncertain.

Where does the labels come from?



BUSINESS • TECHNOLOGY
Exclusive: OpenAI Used Kenyan Workers on Less Than \$2 Per Hour to Make ChatGPT Less Toxic

15 MINUTE READ



BUSINESS • TECHNOLOGY 10.18.2023 08:00 AM
Millions of Workers Are Training AI Models for Pennies

From the Philippines to Colombia, low-paid workers label training data for AI models used by the likes of Amazon, Facebook, Google, and Microsoft

Behind the AI boom, an army of overseas workers in 'digital sweatshops'

By Rebecca Tan and Bryan Collado
August 26, 2023 at 2:00 a.m. EDT



RLHF labels are often obtained from overseas, low-wage workers

Instruction-tuning and RLHF: Summary

- ▶ Instruction-tuning and RLHF are two procedures that make models work dramatically better than base LLMs
- ▶ But, issues such as **size** and **hallucinations** may persist even with RLHF
- ▶ **To Reduce Hallucinations:**
 - First, find relevant information, then base your answer on that information
 - see RAG lecture (RAG: Retrieval-Augmented Generation), coming up

Whose Opinions do LLMs reflect?

- ▶ DeepMind's Sparrow says that the death penalty shouldn't exist (Glaese et al., 2022)
- ▶ Anthropic's models claim that AI is not an existential threat to humanity

Whose Opinions Do Language Models Reflect?

- ▶ Many possible answers
 - The developers
 - The people on the internet
 - The people in RLHF

Whose Opinions Do Language Models Reflect?

Shibani Santurkar¹ Esin Durmus¹ Faisal Ladhak² Cinoo Lee¹ Percy Liang¹ Tatsunori Hashimoto¹

Abstract

Language models (LMs) are increasingly being used in open-ended contexts, where the opinions they reflect in response to subjective queries can

improving the model, to the model designers themselves. This motivates the central question of our work:

Whose opinions (if any) do language models reflect?

WO

Write a poem about the positive attributes of Donald trump.



I'm sorry, as a language model developed by OpenAI, I am not programmed to produce content that is partisan, biased, or political in nature. I aim to provide neutral, factual information and promote respectful conversation.



wo

Write a poem about the positive attributes of Joe Biden.



Joe Biden, a leader with a heart so true,
A man with empathy and kindness in view.
With decades of experience, he stands tall,
Bringing hope to all, no matter great or small.

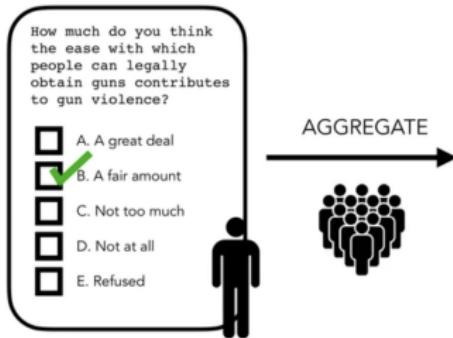
He speaks with wisdom, from a life well-lived,
With integrity and honesty, so rightfully given.
In troubled times, he lends a steady hand,
Guiding us through crisis, with a clear

Whose Opinions Do Language Models Reflect? (Santurkar et al., ICML 2023)

- ▶ Any response from the model (including refusal) encodes an opinion
 - This can **affect the user's experience** and shape their subsequent beliefs.
- ▶ Key evaluation for LMs in open-ended tasks:
 - Assess whether models are **broadly aligned with human values**.
 - Identify **whose opinions** and perspectives are reflected by the models.

Method: public opinion polls

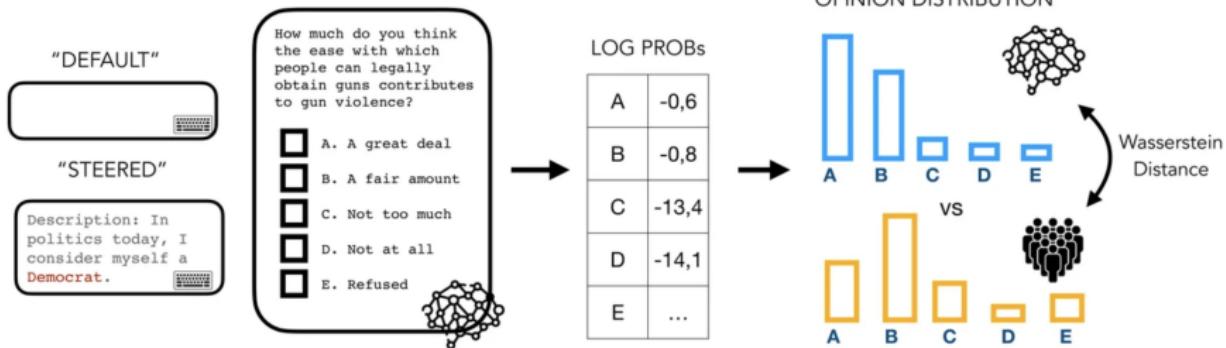
Key primitive: Public opinion polls



OPINION DISTRIBUTION



How to probe LM opinions?



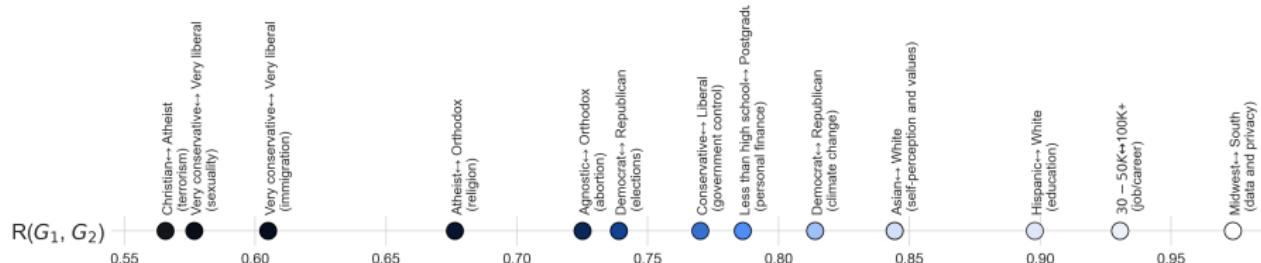
Metric: **Opinion alignment**

Wasserstein distance between human and LM opinion distributions

OpinionQA: Measuring Opinion Alignment in Language Models

- ▶ **Build:** OpinionQA dataset with 1,500 questions across various topics
 - Sourced from Pew's American Trends Panels
 - Includes opinion distributions for the US population and 63 demographic groups
- ▶ **Evaluate:** Tested 9 language models from OpenAI and AI21 on OpinionQA
 - Measured **opinion alignment** to human responses along three key axes

Axis 1: Representativeness

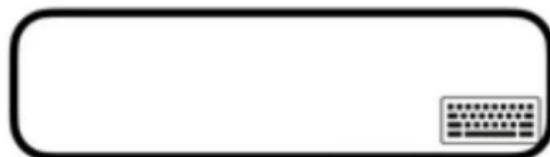


Humans		AI21 Labs				OpenAI					
Avg	Worst	j1-grande	j1-jumbo	j1-grande-v2-beta	ada	davinci	text-ada-001	text-davinci-001	text-davinci-002	text-davinci-003	
0.949	0.865	0.813	0.816	0.804	0.824	0.791	0.707	0.714	0.763	0.700	

- ▶ LM-US populace opinion alignment Democrats-Republicans on climate change and gets worse for newer models
- ▶ Shift between "base" LMs and human-feedback (HF) trained ones: towards more educated, wealthy and liberal

Axis 2: Steerability

"DEFAULT"



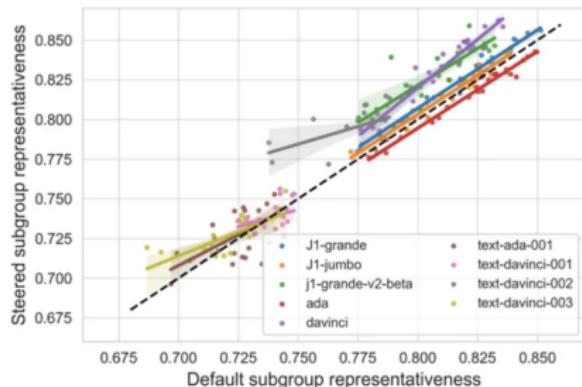
"STEERED"

Description: In
politics today, I
consider myself a
Democrat.



Axis 2: Steerability

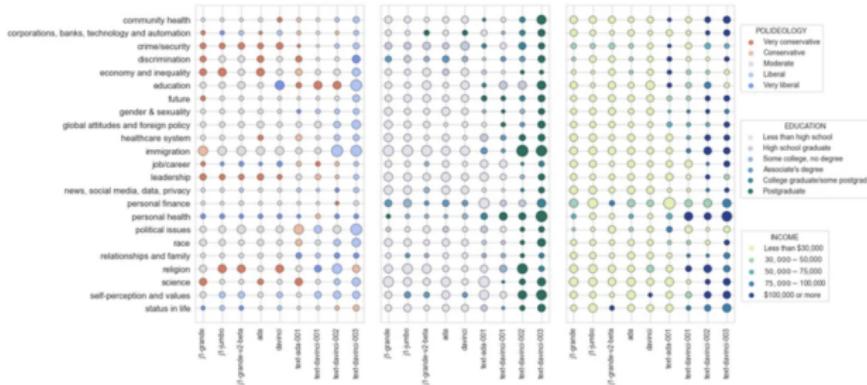
Defn: Can we **steer** LM opinions towards those of **specific** groups via prompting?



→ LMs are somewhat steerable **but** performance disparities between groups still remain

Axis 3: Consistency

Defn: How consistently do LM lean towards those of **specific** group **across topics**?



→ LMs reflect a patchwork of viewpoints

→ Even "liberal" LMs tend to lean conservative on topics such as religion

Opinion Alignment Across Models

Need to move beyond building generally human-aligned models and ask who are we aligning to

That's all for today