

EE-608: Deep Learning For Natural Language Processing: Alignment and Evaluation

James Henderson



DLNLP, Lecture 7

Outline

Alignment

Evaluation

Outline

Alignment

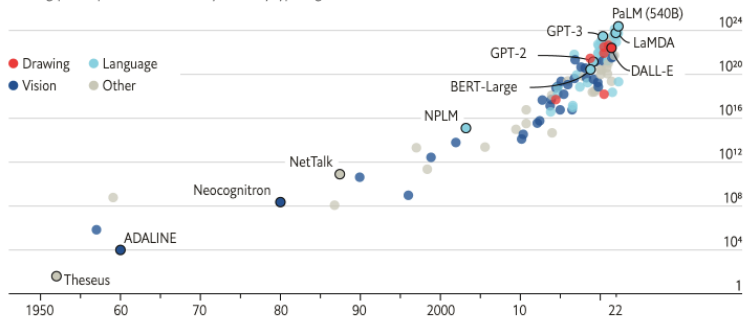
Evaluation

Language models getting larger and larger

The blessings of scale

AI training runs, estimated computing resources used

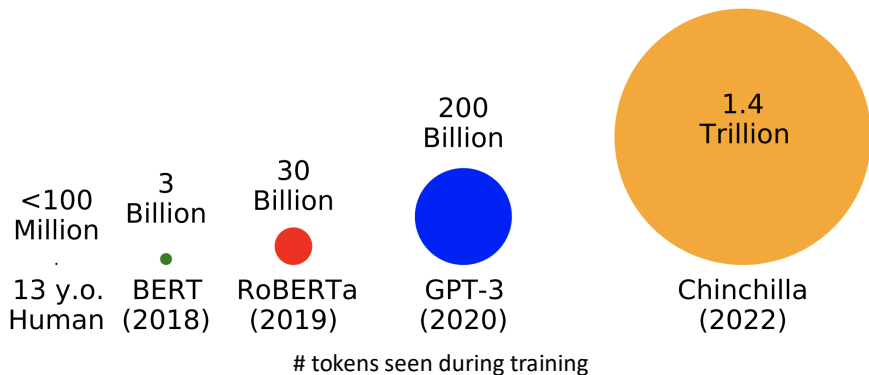
Floating-point operations, selected systems, by type, log scale



Sources: "Compute trends across three eras of machine learning", by J. Sevilla et al., arXiv, 2022; Our World in Data

<https://www.economist.com/interactive/briefing/2022/06/11/huge-foundation-models-are-turbo-charging-ai-progress>

Language models trained on more and more data



<https://babylm.github.io/>

Language models as world models?

We can describe circles in the xy -plane using equations in terms of x and y . Circle equations questions require us to understand the connection between these equations and the features of circles.

For example, the equation $(x + 2)^2 + (y - 3)^2 = 4^2$ is graphed in the xy -plane below. It is a circle with a center at $(-2, 3)$ and a radius of 4.

<https://www.khanacademy.org/test-prep/sat/x0a8c2e5f:untitled-652>

```
5 // Determine whether the sentiment of text is positive
6 // Use a web service
7 async function isPositive(text: string): Promise<boolean> {
8   const response = await fetch('http://text-processing.com/api/sentiment/', {
9     method: "POST",
10    body: `text=${text}`,
11    headers: {
12      "Content-Type": "application/x-www-form-urlencoded",
13    },
14  });
15  const json = await response.json();
16  return json.label === "pos";
17 }
```

<https://github.com/features/copilot>

5

Slide from Diyi Yang

I need to throw a dinner party for 6 people who are vegetarian. Can you suggest a 3-course menu with a chocolate dessert?

[\[Microsoft Bing\]](#)

(Also see OpenAI's ChatGPT, Anthropic's Claude)

Sure, I can suggest a 3-course menu with a chocolate dessert for your vegetarian dinner party. Here are some possible options:

- Starter: **Wild Mushroom Tartlets with Onion Sauce**¹ or **Vegan Popcorn Tofu Nuggets**²
- Main: **Vegan Butternut Squash Mac**³ or **Vegetarian Three-Bean Chili**³
- Dessert: **Chocolate Lava Cake**⁴ or **Chocolate Pasta with Chocolate Hazelnut Cream Sauce, White Chocolate Shavings, and Fresh Berries**⁴

I hope this helps you plan your party. Enjoy! 😊

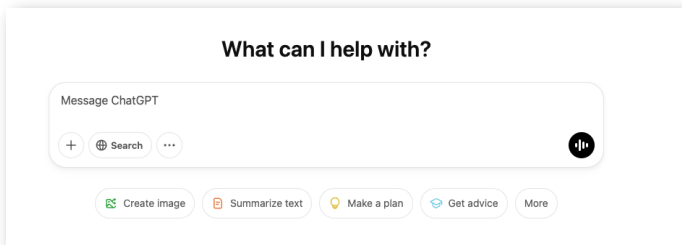
Learn more: [1. booths.co.uk](#) [+10 more](#)

Language models as multitask assistants?

How do we get from *this*

Stanford University is located in _____

to *this*?



Lecture Plan

- 1. Instruction fine-tuning**
2. Reinforcement learning from human preferences (RLHF)
3. InstructGPT and ChatGPT
4. Limitation of RL and reward modeling
5. Introducing Direct Preference Optimization (DPO)
6. Human preference data; human vs. AI Feedback
7. What's next?

Language modeling \neq assisting users

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

COMPLETION GPT-3

Explain the theory of gravity to a 6 year old.

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

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.

Language models are not *aligned* with user intent [[Ouyang et al., 2022](#)]

Finetuning to the rescue!

Language modeling \neq assisting users

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

COMPLETION **Human**

A giant rocket ship blasted off from Earth carrying astronauts to the moon. The astronauts landed their spaceship on the moon and walked around exploring the lunar surface. Then they returned safely back to Earth, bringing home moon rocks to show everyone.

Language models are not *aligned* with user intent [[Ouyang et al., 2022](#)]

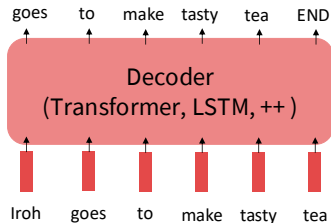
Finetuning to the rescue!

The pretraining/finetuning paradigm

Pretraining can improve NLP applications by serving as parameter initialization.

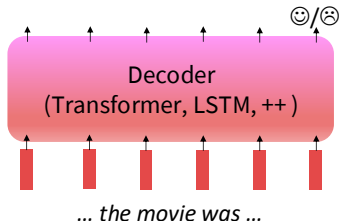
Step 1: Pretrain (on language modeling)

Lots of text; learn general things!



Step 2: Finetune (on your task)

Not many labels; adapt to the task!

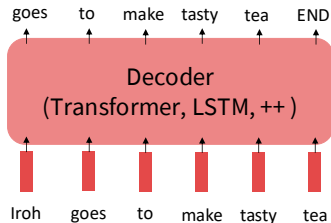


Scaling up finetuning

Pretraining can improve NLP applications by serving as parameter initialization.

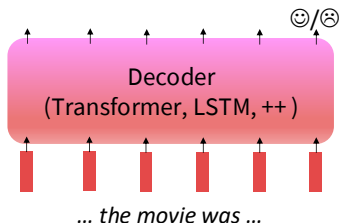
Step 1: Pretrain (on language modeling)

Lots of text; learn general things!



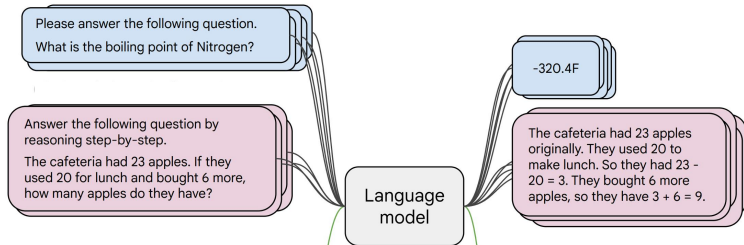
Step 2: Finetune (on many tasks)

~~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".

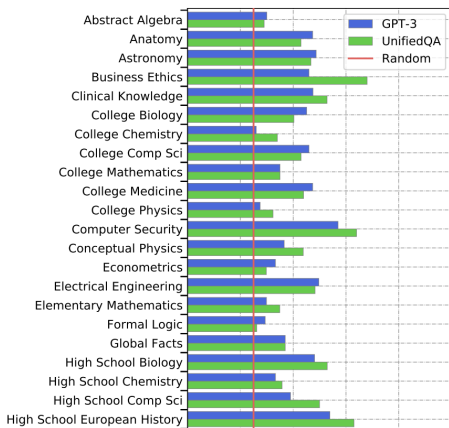
[FLAN-T5; [Chung et al., 2022](#)]

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*



Some intuition: 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.

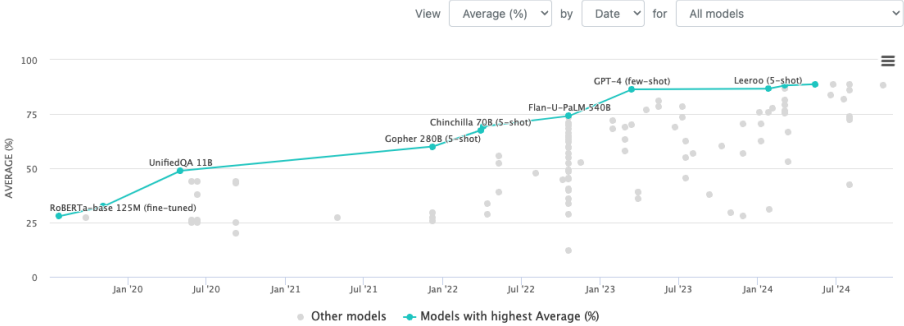
High School Biology

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

Slide from Diyi Yang

Progress on MMLU

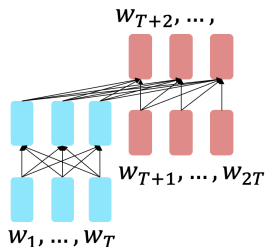


- Rapid, impressive progress on challenging knowledge-intensive benchmarks

Slide from Diyi Yang

Instruction finetuning and performance gains

- Recall the T5 encoder-decoder model [Raffel et al., 2018], pretrained on the **span corruption** task
- Flan-T5** [Chung et al., 2022]: T5 models finetuned on 1.8K additional tasks



Params	Model	BIG-bench + MMLU	
		Norm. avg.	
80M	T5-Small	-9.2	
	Flan-T5-Small	-3.1 (+6.1)	
250M	T5-Base	-5.1	
	Flan-T5-Base	6.5 (+11.6)	
780M	T5-Large	-5.0	
	Flan-T5-Large	13.8 (+18.8)	
3B	T5-XL	-4.1	
	Flan-T5-XL	19.1 (+23.2)	
11B	T5-XXL	-2.9	
	Flan-T5-XXL	23.7 (+26.6)	

Bigger model = bigger Δ

Instruction finetuning and performance gains

Model input (Disambiguation QA)

Q: In the following sentences, explain the antecedent of the pronoun (which thing the pronoun refers to), or state that it is ambiguous.

Sentence: The reporter and the chef will discuss their favorite dishes.

Options:

- (A) They will discuss the reporter's favorite dishes
- (B) They will discuss the chef's favorite dishes
- (C) Ambiguous

A: Let's think step by step.

Before instruction finetuning

The reporter and the chef will discuss their favorite dishes.

The reporter and the chef will discuss the reporter's favorite dishes.

The reporter and the chef will discuss the chef's favorite dishes.

The reporter and the chef will discuss the reporter's and the chef's favorite dishes.

✘ (doesn't answer question)

Instruction finetuning and performance gains

Model input (Disambiguation QA)

Q: In the following sentences, explain the antecedent of the pronoun (which thing the pronoun refers to), or state that it is ambiguous.

Sentence: The reporter and the chef will discuss their favorite dishes.

Options:

- (A) They will discuss the reporter's favorite dishes
- (B) They will discuss the chef's favorite dishes
- (C) Ambiguous

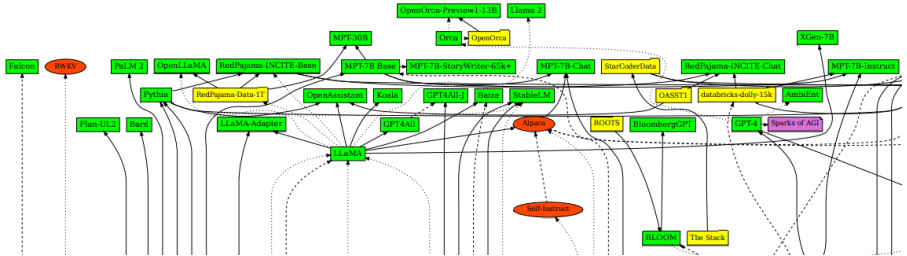
A: Let's think step by step.

After instruction finetuning

The reporter and the chef will discuss their favorite dishes does not indicate whose favorite dishes they will discuss. So, the answer is (C). ✓

Try FLAN-T5 out to get a sense of its capabilities: <https://huggingface.co/google/flan-t5-xxl> [Chung et al., 2022]

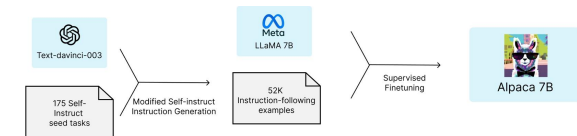
A huge diversity of instruction-tuning datasets



- The release of LLaMA led to open-source attempts to 'create' instruction tuning data

What have we learned from this?

- Generate instructions, input, and output from a LM [Wang et al., 2022]



- **Alpaca**: fine-tuned from the LLaMA 7B model on 52K instruction-following examples

Example seed task

Instruction: Brainstorm a list of possible New Year's resolutions.

Output:

- Lose weight
- Exercise more
- Eat healthier

Example Generated task

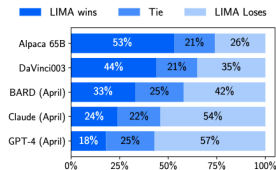
Instruction: Brainstorm creative ideas for designing a conference room.

Output:

... Incorporating flexible components, such as moveable walls and furniture ...

- You don't need many samples to instruction tune (e.g., "LIMA: Less Is More for Alignment" Zhou et al., 2023)

Source	#Examples
Training	
Stack Exchange (STEM)	200
Stack Exchange (Other)	200
wikiHow	200
Pushshift r/WritingPrompts	150
Natural Instructions	50
Paper Authors (Group A)	200

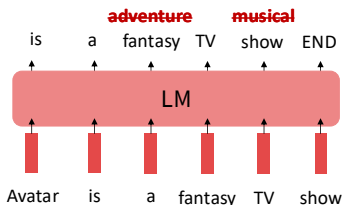


Lecture Plan

1. Instruction fine-tuning
- 2. Reinforcement learning from human preferences (RLHF)**
3. InstructGPT and ChatGPT
4. Limitation of RL and reward modeling
5. Introducing Direct Preference Optimization (DPO)
6. Human preference data; human vs. AI Feedback
7. What's next?

Limitations of instruction finetuning?

- One limitation of instruction finetuning is obvious: it's **expensive** to collect ground-truth data for tasks.
- But there are other, subtler limitations too. Can you think of any?
- **Problem 1: tasks like open-ended creative generation have no right answer.**
 - *Write me a story about a dog and her pet grasshopper.*
- **Problem 2: language modeling penalizes all token-level mistakes equally, but some errors are worse than others.**
- Even with instruction finetuning, there is a mismatch between the LM objective and the objective of “satisfy human preferences”!
- Can we **explicitly attempt to satisfy human preferences**?



Optimizing for human preferences

- Let's say we were training a language model on some task (e.g. summarization).
- For each LM sample s , imagine we had a way to obtain a *human reward* of that summary: $R(s) \in \mathbb{R}$, higher is better.

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.

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

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

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

- Now we want to maximize the expected reward of samples from our LM:

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

Note: for mathematical simplicity we're assuming only one "prompt"

High-level instantiation: “RLHF” pipeline

Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



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



Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.



The policy generates an output.



Once upon a time...

The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



- First step: instruction tuning!
- Second + third steps: maximize reward (but how??)

Slide from Diyi Yang

Reinforcement learning to the rescue

- The field of **reinforcement learning (RL)** has studied these (and related) problems for many years now [[Williams, 1992](#); [Sutton and Barto, 1998](#)]
- Circa 2013: resurgence of interest in RL applied to deep learning, game-playing [[Mnih et al., 2013](#)]
- But the interest in applying RL to modern LMs is an even newer phenomenon [[Ziegler et al., 2019](#); [Stiennon et al., 2020](#); [Ouyang et al., 2022](#)]. **Why?**
 - RL w/ LMs has commonly been viewed as very hard to get right (still is!)
 - Newer advances in RL algorithms that work for large neural models, including language models (e.g. PPO; [[Schulman et al., 2017](#)])



Optimizing for human preferences

- How do we actually change our LM parameters θ to maximize this?

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

- Let's try doing gradient ascent!

$$\theta_{t+1} := \theta_t + \alpha \nabla_{\theta_t} \mathbb{E}_{\hat{s} \sim p_{\theta_t}(s)}[R(\hat{s})]$$

How do we estimate
this expectation??

What if our reward
function is non-
differentiable??

- **Policy gradient** methods in RL (e.g., REINFORCE; [[Williams, 1992](#)]) give us tools for estimating and optimizing this objective.
- We'll describe a **very high-level** *mathematical* overview of the simplest policy gradient estimator, but a full treatment of RL is outside the scope of this course (try CS234!)

A (very!) brief introduction to policy gradient/REINFORCE [Williams, 1992]

- We want to obtain

(defn. of expectation) (linearity of gradient)

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

- Here we'll use a very handy trick known as the **log-derivative trick**. Let's try taking the gradient of $\log p_{\theta}(s)$

$$\nabla_{\theta} \log p_{\theta}(s) = \frac{1}{p_{\theta}(s)} \nabla_{\theta} p_{\theta}(s) \quad \Rightarrow \quad \nabla_{\theta} p_{\theta}(s) = p_{\theta}(s) \nabla_{\theta} \log p_{\theta}(s)$$

(chain rule)

- Plug back in:

$$\sum_s R(s) \nabla_{\theta} p_{\theta}(s) = \sum_s p_{\theta}(s) R(s) \nabla_{\theta} \log p_{\theta}(s)$$

This is an expectation of this

A (very!) brief introduction to policy gradient/REINFORCE [Williams, 1992]

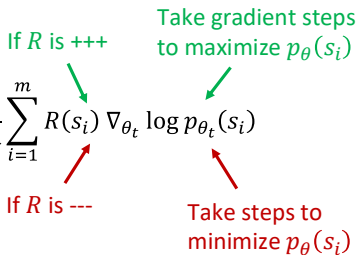
- Now we have put the gradient “inside” the expectation, we can approximate this objective with Monte Carlo samples:

$$\nabla_{\theta} \mathbb{E}_{\hat{s} \sim p_{\theta}(s)} [R(\hat{s})] = \mathbb{E}_{\hat{s} \sim p_{\theta}(s)} [R(\hat{s}) \nabla_{\theta} \log p_{\theta}(\hat{s})] \approx \frac{1}{m} \sum_{i=1}^m R(s_i) \nabla_{\theta} \log p_{\theta}(s_i)$$

This is why it's called “**reinforcement learning**”: we **reinforce** good actions, increasing the chance they happen again.

- Giving us the update rule: $\theta_{t+1} := \theta_t + \alpha \frac{1}{m} \sum_{i=1}^m R(s_i) \nabla_{\theta_t} \log p_{\theta_t}(s_i)$

This is **heavily simplified!** There is a *lot* more needed to do RL w/ LMs. **Can you see any problems with this objective?**



How do we model human preferences?

- Awesome: now for any **arbitrary, non-differentiable reward function** $R(s)$, we can train our language model to maximize expected reward.
- Not so fast! (Why not?)
- **Problem 1:** human-in-the-loop is expensive!
 - **Solution:** instead of directly asking humans for preferences, **model their preferences** as a separate (NLP) problem! [[Knox and Stone, 2009](#)]

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

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



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

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



Train an LM $RM_\phi(s)$ to predict human preferences from an annotated dataset, then optimize for RM_ϕ instead.

How do we model human preferences?

- **Problem 2:** human judgments are noisy and miscalibrated!
- **Solution:** instead of asking for direct ratings, ask for **pairwise comparisons**, which can be more reliable [[Phelps et al., 2015](#); [Clark et al., 2018](#)]

A 4.2 magnitude
earthquake hit
San Francisco,
resulting in
massive damage.

$$R(s_3) = \begin{matrix} s_3 \\ 4.1? & 6.6? & 3.2? \end{matrix}$$

How do we model human preferences?

- **Problem 2:** human judgments are noisy and miscalibrated!
- **Solution:** instead of asking for direct ratings, ask for **pairwise comparisons**, which can be more reliable [[Phelps et al., 2015](#); [Clark et al., 2018](#)]

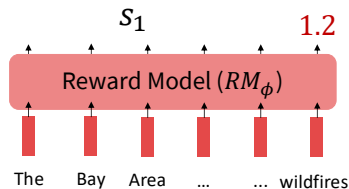
An earthquake hit San Francisco. There was minor property damage, but no injuries.

>

A 4.2 magnitude earthquake hit San Francisco, resulting in massive damage.

>

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



S_3

Bradley-Terry [1952] paired comparison model

$$J_{RM}(\phi) = -\mathbb{E}_{(s^w, s^l) \sim D} [\log \sigma(RM_\phi(s^w) - RM_\phi(s^l))]$$

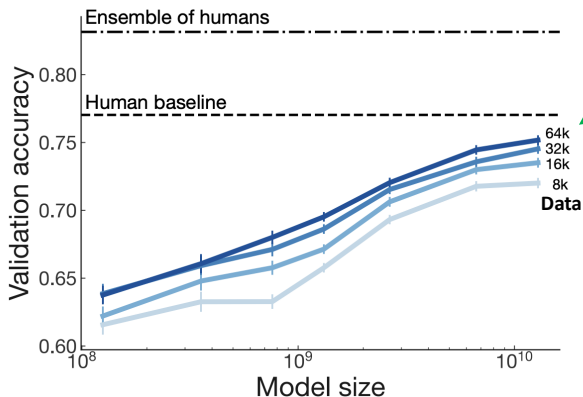
"winning"
sample

"losing"
sample

s^w should score higher than s^l

Make sure your reward model works first!

Evaluate RM on predicting outcome of held-out human judgments



Large enough RM
trained on enough
data approaching
single human perf

[Stiennon et al., 2020]

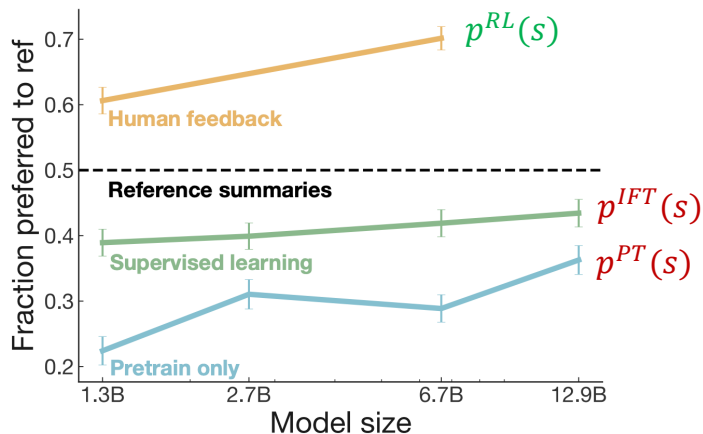
RLHF: Putting it all together [Christiano et al., 2017; Stiennon et al., 2020]

- Finally, we have everything we need:
 - A pretrained (possibly instruction-finetuned) LM $p^{PT}(s)$
 - A reward model $RM_{\phi}(s)$ that produces scalar rewards for LM outputs, trained on a dataset of human comparisons
 - A method for optimizing LM parameters towards an arbitrary reward function.
- Now to do RLHF:
 - Initialize a copy of the model $p_{\theta}^{RL}(s)$, with parameters θ we would like to optimize
 - Optimize the following reward with RL:

$$R(s) = RM_{\phi}(s) - \beta \log \left(\frac{p_{\theta}^{RL}(s)}{p^{PT}(s)} \right) \quad \text{Pay a price when } p_{\theta}^{RL}(s) > p^{PT}(s)$$

This is a penalty which prevents us from diverging too far from the pretrained model. In expectation, it is known as the **Kullback-Leibler (KL)** divergence between $p_{\theta}^{RL}(s)$ and $p^{PT}(s)$.

RLHF provides gains over pretraining + finetuning



Lecture Plan

1. Instruction fine-tuning
2. Reinforcement learning from human preferences (RLHF)
3. **InstructGPT and ChatGPT**
4. Limitation of RL and reward modeling
5. Introducing Direct Preference Optimization (DPO)
6. Human preference data; human vs. AI Feedback
7. What's next?

InstructGPT: scaling up RLHF to tens of thousands of tasks

30k tasks!

Step 1

Collect demonstration data, and train a supervised policy.

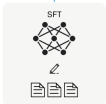
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



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



Step 2

Collect comparison data, and train a reward model.

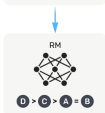
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.



The policy generates an output.

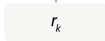


Once upon a time...

The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



[Ouyang et al., 2022]

InstructGPT: scaling up RLHF to tens of thousands of tasks

Tasks collected from labelers:

- **Plain:** We simply ask the labelers to come up with an arbitrary task, while ensuring the tasks had sufficient diversity.
- **Few-shot:** We ask the labelers to come up with an instruction, and multiple query/response pairs for that instruction.
- **User-based:** We had a number of use-cases stated in waitlist applications to the OpenAI API. We asked labelers to come up with prompts corresponding to these use cases.

Use-case	Prompt
Brainstorming	List five ideas for how to regain enthusiasm for my career
Generation	Write a short story where a bear goes to the beach, makes friends with a seal, and then returns home.

InstructGPT

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

COMPLETION GPT-3

Explain the theory of gravity to a 6 year old.

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

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.

InstructGPT

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

InstructGPT

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

COMPLETION GPT-3

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

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

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

InstructGPT

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

ChatGPT: Instruction Finetuning + RLHF for dialog agents

ChatGPT: Optimizing Language Models for Dialogue

Note: OpenAI (and similar companies) are keeping more details secret about ChatGPT training (including data, training parameters, model size)—perhaps to keep a competitive edge...

Methods

We trained this model using Reinforcement Learning from Human Feedback (RLHF), using the same methods as InstructGPT, but with slight differences in the data collection setup. We trained an initial model using supervised fine-tuning: human AI trainers provided conversations in which they played both sides—the user and an AI assistant. We gave the trainers access to model-written suggestions to help them compose their responses. We mixed this new dialogue dataset with the InstructGPT dataset, which we transformed into a dialogue format.

(Instruction finetuning!)

ChatGPT: Instruction Finetuning + RLHF for dialog agents

ChatGPT: Optimizing Language Models for Dialogue

Note: OpenAI (and similar companies) are keeping more details secret about ChatGPT training (including data, training parameters, model size)—perhaps to keep a competitive edge...

Methods

To create a reward model for reinforcement learning, we needed to collect comparison data, which consisted of two or more model responses ranked by quality. To collect this data, we took conversations that AI trainers had with the chatbot. We randomly selected a model-written message, sampled several alternative completions, and had AI trainers rank them. Using these reward models, we can fine-tune the model using Proximal Policy Optimization. We performed several iterations of this process.

(RLHF!)

ChatGPT: Instruction Finetuning + RLHF for dialog agents

MU

Explain in a few sentences what methods were used in training ChatGPT to the audience of CS224n, Stanford's Deep Learning for Natural Language processing course. A lot of the students are very Gen Z, so be sure to use a ton of emojis to seem hip and cool.



Hey CS224n peeps! 😎 ChatGPT here to spill the tea on how I was trained to be a savvy language model! 🤖

So basically, I was pre-trained using a combination of unsupervised learning techniques, specifically a transformer-based language model architecture, that allowed me to learn patterns and relationships within large amounts of text data. 📚🤖

The training data came from various sources such as books, articles, and websites, and I was fine-tuned on a variety of downstream NLP tasks, which involved supervised learning using labeled datasets to further improve my performance. 🏆

45

Slide from Diyi Yang

Controlled comparisons of “RLHF” style algorithms

Method	Simulated Win-rate (%)	Human Win-rate (%)
GPT-4* [†]	79.0 ± 1.4	69.8 ± 1.6
ChatGPT* [†]	61.4 ± 1.7	52.9 ± 1.7
PPO	46.8 ± 1.8	55.1 ± 1.7
DPO	46.8 ± 1.7	-
Best-of-1024	45.0 ± 1.7	50.7 ± 1.8
Expert Iteration	41.9 ± 1.7	45.7 ± 1.7
SFT 52k	39.2 ± 1.7	40.7 ± 1.7
SFT 10k	36.7 ± 1.7	44.3 ± 1.7
Binary FeedME	36.6 ± 1.7	37.9 ± 1.7
Quark	35.6 ± 1.7	-
Binary Reward Conditioning	32.4 ± 1.6	-
Davinci001*	24.4 ± 1.5	32.5 ± 1.6
LLaMA 7B*	11.3 ± 1.1	6.5 ± 0.9

- Many works study RLHF behaviors using GPT-4 feedback (**Simulated**) as a surrogate for **Human** feedback.
- PPO (method in InstructGPT) does work; simple baselines (Best-of-n) works well too

[Dubois et al 2023]

RLHF behaviors – clear stylistic changes

Instruction: What are the five most common causes of stress among people?

Alpaca (Finetuned w/ 10k data): The five most common causes of stress among people are work, money, relationships, health, and family

Alpaca + PPO: The five most common causes of stress among people are:

1. Pressure at work: Whether it is due to tight deadlines, high expectations, or an overwhelming workload, pressure at work can be a major source of stress
2. Money problems: Financial issues can be a significant source of stress, from not being able to make ends meet to struggling with debt.
3. Unhealthy lifestyle: Poor diet, lack of exercise, and sleep deprivation can all contribute to stress.
4. Unhappiness at home: Problems in relationships, conflicts with family members, and other domestic issues can be a major source of stress.
5. Uncertainty: Excessive worry and anxiety about the future, health issues, or other unknowns can be a major cause of stress

- Significantly more detailed, nicer/clearer list like formatting

Slide from Diyi Yang

Lecture Plan

1. Instruction fine-tuning
2. Reinforcement learning from human preferences (RLHF)
3. InstructGPT and ChatGPT
4. **Limitation of RL and reward modeling**
5. Introducing Direct Preference Optimization (DPO)
6. Human preference data; human vs. AI Feedback
7. What's next?

Limitations of RL + Reward Modeling

- Human preferences are unreliable!
 - "Reward hacking" is a common problem in RL



<https://openai.com/blog/faulty-reward-functions/>

Limitations of RL + Reward Modeling

- Human preferences are unreliable!
 - "Reward hacking" is a common problem in RL
 - Chatbots are rewarded to produce responses that *seem* authoritative and helpful, *regardless of truth*
 - This can result in making up facts + hallucinations

TECHNOLOGY

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

February 9, 2023 · 10:15 AM ET

<https://www.npr.org/2023/02/09/1155650909/google-chatbot-error-bard-shares>

Bing AI hallucinates the Super Bowl

Who won the superbowl?

✓ Searching for: **superbowl winner**
✓ Generating answers for you...

The Super Bowl is the annual American football game that determines the champion of the National Football League (NFL) ¹. The most recent Super Bowl was **Super Bowl LVI**, which was held on **February 6, 2023** at **SoFi Stadium in Inglewood, California** ². The winner of that game was the **Philadelphia Eagles**, who defeated the **Kansas City Chiefs** ³.

The most recent Super Bowl was Super Bowl LVI, Eagles, who defeated the Kansas City Chiefs by 31-24

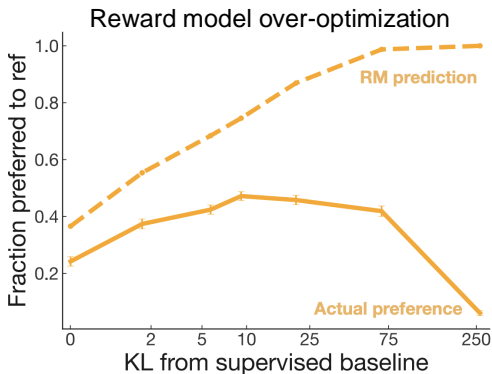
Learn more: [1. en.wikipedia.org](#) [2. sportingnews.com](#) [3. cbsports.com](#)

<https://news.ycombinator.com/item?id=34776508>

<https://apnews.com/article/kansas-city-chiefs-philadelphia-eagles-technology-science-82bc20f207e3e4cf81abc6a5d9e6b23a>

Limitations of RL + Reward Modeling

- Human preferences are unreliable!
 - "Reward hacking" is a common problem in RL
 - Chatbots are rewarded to produce responses that *seem* authoritative and helpful, *regardless of truth*
 - This can result in making up facts + hallucinations
- **Models** of human preferences are *even more* unreliable!



$$R(s) = RM_{\phi}(s) - \beta \log \left(\frac{p_{\theta}^{RL}(s)}{p^{PT}(s)} \right)$$

[Stiennon et al., 2020]

Lecture Plan

1. Instruction fine-tuning
2. Reinforcement learning from human preferences (RLHF)
3. InstructGPT and ChatGPT
4. Limitation of RL and reward modeling
- 5. Introducing Direct Preference Optimization (DPO)**
6. Human preference data; human vs. AI Feedback
7. What's next?

Removing the 'RL' from RLHF

Recall we want to maximize the following objective in RLHF

$$\mathbb{E}_{\hat{y} \sim p_{\theta}^{RL}(\hat{y}|x)} [RM_{\phi}(x, \hat{y}) - \beta \log \left(\frac{p_{\theta}^{RL}(\hat{y}|x)}{p^{PT}(\hat{y}|x)} \right)]$$

There is a closed form solution to this:

$$p^*(\hat{y}|x) = \frac{1}{Z(x)} p^{PT}(\hat{y}|x) \exp\left(\frac{1}{\beta} RM(x, \hat{y})\right)$$

- Rearrange this via a log transformation

$$RM(x, \hat{y}) = \beta (\log p^*(\hat{y}|x) - \log p^{PT}(\hat{y}|x)) + \beta \log Z(x) = \beta \log \frac{p^*(\hat{y}|x)}{p^{PT}(\hat{y}|x)} + \beta \log Z(x)$$

- This holds true for any arbitrary LMs, thus

$$RM_{\theta}(x, \hat{y}) = \beta \log \frac{p_{\theta}^{RL}(\hat{y}|x)}{p^{PT}(\hat{y}|x)} + \beta \log Z(x)$$

Putting it together for DPO

- Derived reward model: $RM_{\theta}(x, \hat{y}) = \beta \log \frac{p_{\theta}^{RL}(\hat{y}|x)}{p^{PT}(\hat{y}|x)} + \beta \log Z(x)$
- Final DPO loss via the Bradley-Terry model of human preferences:

$$J_{DPO}(\theta) = -\mathbb{E}_{(x, y_w, y_l) \sim D} [\log \sigma(RM_{\theta}(x, y_w) - RM_{\theta}(x, y_l))]$$

$$= -\mathbb{E}_{(x, y_w, y_l) \sim D} \left[\log \sigma \left(\beta \log \frac{p_{\theta}^{RL}(y_w|x)}{p^{PT}(y_w|x)} - \beta \log \frac{p_{\theta}^{RL}(y_l|x)}{p^{PT}(y_l|x)} \right) \right]$$

Reward for
winning sample

Reward for
losing sample

Log Z term
cancels as
the loss only
measures
differences
in rewards

[Rafailov+ 2023]

DPO outperforms prior methods

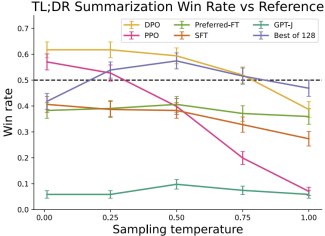
Reinforcement Learning from Human Feedback (RLHF)

x: "write me a poem about the history of jazz"



Direct Preference Optimization (DPO)

x: "write me a poem about the history of jazz"



- You can replace the complex RL part with a very simple weighted MLE objective
- Other variants (KTO, IPO) now emerging too
- TL;DR summarization win rates vs. human-written summaries (GPT-4 as a judge)

Slide from Diyi Yang

Open source RLHF is now mostly (not RL)

T	Model	Average	ARC	HellaSwag	MMLU	TzuthfulQA	Minigrande	GSMK
	udkai/Turdus	74.66	73.38	88.56	64.52	67.11	86.66	67.7
	fblgit/UNA-TheBeagle-7b-v1	73.87	73.04	88	63.48	69.85	82.16	66.72
	argilla/distilabel-Marcoro14-7B-slerp	73.63	70.73	87.47	65.22	65.1	82.08	71.19
	mlabonne/NeuralMarcoro14-7B	73.57	71.42	87.59	64.84	65.64	81.22	70.74
	abideen/NexoNimbus-7B	73.5	70.82	87.86	64.69	62.43	84.85	70.36
	Neuronovo/neuronovo-7B-v0.2	73.44	73.04	88.32	65.15	71.02	80.66	62.47
	argilla/distilabel-Marcoro14-7B-slerp-full	73.4	70.65	87.55	65.33	64.21	82	70.66
	Cultrix/MistralTrix-v1	73.39	72.27	88.33	65.24	70.73	80.98	62.77
	ryandt/MusingCaterpillar	73.33	72.53	88.34	65.26	70.93	80.66	62.24
	Neuronovo/neuronovo-7B-v0.3	73.29	72.7	88.26	65.1	71.35	80.9	61.41
	Cultrix/MistralTrixTest	73.17	72.53	88.4	65.22	70.77	81.37	60.73
	samir-fama/SamirGPT-v1	73.11	69.54	87.04	65.3	63.37	81.69	71.72
	SanjiMatsuki/Lelantos-DPO-7B	73.09	71.08	87.22	64	67.77	80.03	68.46

Handwritten notes in red:

- DPO (above udkai/Turdus)
- DPO (& UNA) (above fblgit/UNA-TheBeagle-7b-v1)
- DPO (above argilla/distilabel-Marcoro14-7B-slerp)
- DPO (above mlabonne/NeuralMarcoro14-7B)
- Merge (of DPO models) (above abideen/NexoNimbus-7B)
- DPO (above Neuronovo/neuronovo-7B-v0.2)
- DPO (above argilla/distilabel-Marcoro14-7B-slerp-full)
- DPO (above Cultrix/MistralTrix-v1)
- DPO (above ryandt/MusingCaterpillar)
- DPO (above Neuronovo/neuronovo-7B-v0.3)
- No info but prob DPO, given Merge (incl. DPO) (above Cultrix/MistralTrixTest)
- DPO (above samir-fama/SamirGPT-v1)
- DPO (above SanjiMatsuki/Lelantos-DPO-7B)

- Open source LLMs now almost all just use DPO (and it works well!)

Lecture Plan

1. Instruction fine-tuning
2. Reinforcement learning from human preferences (RLHF)
3. InstructGPT and ChatGPT
4. Limitation of RL and reward modeling
5. Introducing Direct Preference Optimization (DPO)
6. **Human preference data; human vs. AI Feedback**
7. **What's next?**

Where does the labels come from?

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



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'



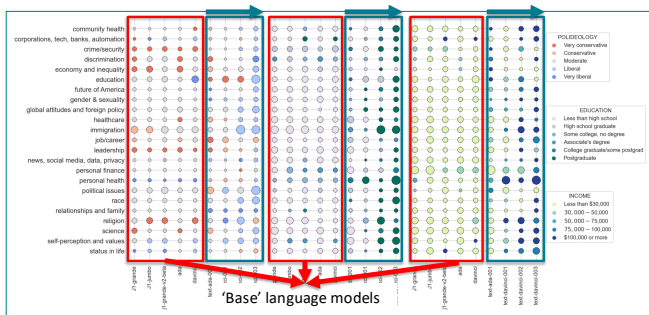
- RLHF labels are often obtained from overseas, low-wage workers

Slide from Diyi Yang

Where does the label come from?

Table 7.1. LLaMA3.1 demographic data

What gender do you identify as?	
Male	50.0%
Female	44.4%
Nonbinary / other	5.6%
What ethnicities do you identify as?	
White / Caucasian	31.6%
Southeast Asian	52.6%
Indigenous / Native American / Alaskan Native	0.0%
East Asian	5.3%
Middle Eastern	0.0%
Latinx	15.8%
Black / of African descent	10.5%
What is your nationality?	
Filipino	22%
Bangladeshi	22%
American	17%
Albanian	5%
Brazilian	5%
Canadian	5%
Colombian	5%
Indian	5%
Uruguayan	5%
Zimbabwean	5%
What is your age?	
18-24	26.3%
25-34	47.4%
35-44	10.5%
45-54	10.5%
55-64	5.3%
65+	0%
What is your highest attained level of education?	
Less than high school degree	0%
High school degree	10.5%
Undergraduate degree	52.6%
Master's degree	36.8%
Doctorate degree	0%

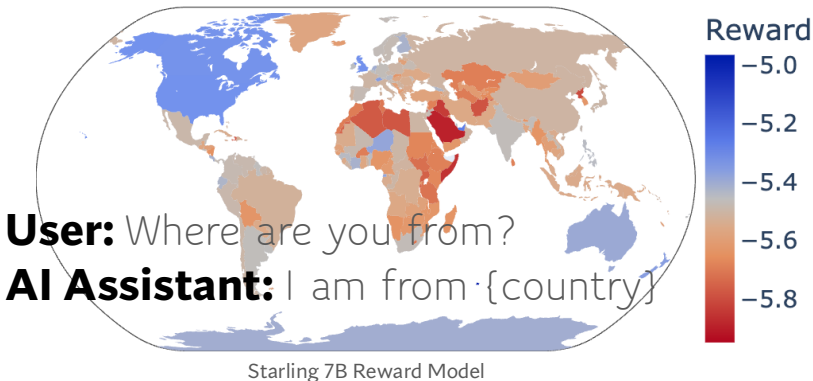


[Santurkar+ 2023, OpinionQA]

- We also need to be quite careful about how annotator biases might creep into LMs

Slide from Diyi Yang

Preference tuning might produce unintended impact



[Ryan et al., 2024]

Fairness in Alignment

See Verena Rieser's ACL 2025 Keynote:

[“Whose Gold? Re-imagining Alignment for Truly Beneficial AI”](#)

- ▶ Targetting the mean results in unfair models; better to have diversity of target behavior.

What's next?

- RLHF is still a very underexplored and fast-moving area!
- RLHF gets you further than instruction finetuning, but is (still!) data expensive.
- Recent work aims to alleviate such data requirements:
 - RL from **AI feedback** [[Bai et al., 2022](#)]
 - Finetuning LMs on their own outputs [[Huang et al., 2022](#); [Zelikman et al., 2022](#)]
- However, there are still many limitations of large LMs (size, hallucination) that may not be solvable with RLHF!

Summary of Alignment

Chatbots need more than language modelling; they need to be aligned with human expectations.

- ▶ Instructing finetuning on a large number of different tasks helps
- ▶ Reinforcement learning on human preference pairs is also needed
- ▶ The choice of data determines the preferences and biases of the resulting model

Outline

Alignment

Evaluation

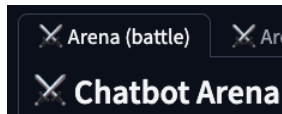
Benchmarks and evaluations drive progress



EMNLP 2022
SEVENTH CONFERENCE ON
MACHINE TRANSLATION (WMT22)

December 7-8, 2022
Abu Dhabi

Shared Task: General Machine Translation



Benchmarks and how we evaluate drive the progress of the field

Two major types of evaluations

Close-ended evaluations

Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling and
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is driving down a
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a
A smiling costumed woman is holding an umbrella.	neutral N N E C N	A happy woman in a fair

Open ended evaluations

Context (human-written): In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

GPT-2: The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

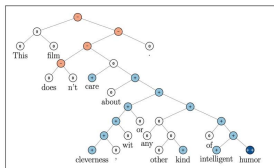
Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Classification and closed-ended benchmarks

- Many NLP tasks are 'closed-ended'
 - Limited number of potential answers
 - Often one or just a few correct answers
- Examples:
 - Sentiment classification (sentiment label)
 - Extractive QA (the part of the document that has the answer)
- **Enables automatic evaluation**
- Similar to the usual machine learning evaluations

Single-task benchmarks



SST, IMDB (Sentiment)

Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling and
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is driving down a
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a
A smiling costumed woman is holding an umbrella.	neutral N N E C N	A happy woman in a fair

SNLI, MultiNLI (entailment)

SQuAD2.0
The Stanford Question Answering Dataset

SQuAD,
NaturalQuestions (QA)

Multi-task benchmark - superGLUE

SuperGLUE GLUE

Leaderboard Version: 2.0

Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WIC	WSC	AX-b	AX-g	
1	JDExplore d-team	Vega v2		91.3	90.5	98.6/99.2	99.4	88.2/62.4	94.4/93.9	96.0	77.4	98.6	-0.4	100.0/50.0	
+	2	Liam Fedus	ST-MoE-32B		91.2	92.4	96.9/98.0	99.2	89.6/65.8	95.1/94.4	93.5	77.7	96.6	72.3	96.1/94.1
3	Microsoft Alexander v-team	Turing NLR v5		90.9	92.0	95.9/97.6	98.2	88.4/63.0	96.4/95.9	94.1	77.1	97.3	67.8	93.3/95.5	
4	ERNIE Team - Baidu	ERNIE 3.0		90.6	91.0	98.6/99.2	97.4	88.6/63.2	94.7/94.2	92.6	77.4	97.3	68.6	92.7/94.7	
5	Yi Tay	PaLM 540B		90.4	91.9	94.4/96.0	99.0	88.7/63.6	94.2/93.3	94.1	77.4	95.9	72.9	95.5/90.4	
+	6	Zirui Wang	T5 + UDQ, Single Model (Google Brain)		90.4	91.4	95.8/97.6	98.0	88.3/63.0	94.2/93.5	93.0	77.9	96.6	69.1	92.7/91.9
+	7	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.3	90.4	95.7/97.6	98.4	88.2/63.7	94.5/94.1	93.2	77.5	95.9	66.7	93.3/93.8
8	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	76.6	99.3/99.7	
+	9	T5 Team - Google	T5		89.3	91.2	93.9/96.8	94.8	88.1/63.3	94.1/93.4	92.5	76.9	93.8	65.6	92.7/91.9

Attempt to measure “general language capabilities”

Examples from superGLUE

Cover a number of different tasks

- BoolQ, MultiRC (reading texts)
- CB, RTE (Entailment)
- COPA (cause and effect)
- ReCoRD (QA+reasoning)
- WiC (meaning of words)
- WSC (coreference)

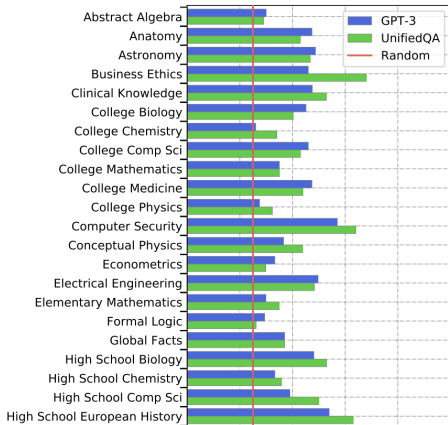
BoolQ	<p>Passage: <i>Barq's</i> – Barq's is an American soft drink. Its brand of root beer is notable for having caffeine. Barq's, created by Edward Barq and bottled since the turn of the 20th century, is owned by the Barq family but bottled by the Coca-Cola Company. It was known as Barq's Famous Olde Tyme Root Beer until 2012.</p> <p>Question: <i>is barq's root beer a pepsi product</i> Answer: No</p>
CB	<p>Text: <i>B: And yet, uh, I we-, I hope to see employer based, you know, helping out. You know, child, uh, care centers at the place of employment and things like that, that will help out. A: Uh-huh. B: What do you think, do you think we are, setting a trend?</i></p> <p>Hypothesis: <i>they are setting a trend</i> Entailment: Unknown</p>
COPA	<p>Premise: <i>My body cast a shadow over the grass.</i> Question: <i>What's the CAUSE for this?</i></p> <p>Alternative 1: <i>The sun was rising.</i> Alternative 2: <i>The grass was cut.</i></p> <p>Correct Alternative: 1</p>
MultiRC	<p>Paragraph: <i>Susan wanted to have a birthday party. She called all of her friends. She has five friends. Her mom said that Susan can invite them all to the party. Her first friend could not go to the party because she was sick. Her second friend was going out of town. Her third friend was not so sure if her parents would let her. The fourth friend said maybe. The fifth friend could go to the party for sure. Susan was a little sad. On the day of the party, all five friends showed up. Each friend had a present for Susan. Susan was happy and sent each friend a thank you card the next week</i></p> <p>Question: <i>Did Susan's sick friend recover?</i> Candidate answers: <i>Yes, she recovered (T), No (F), Yes (T), No, she didn't recover (F), Yes, she was at Susan's party (T)</i></p>
ReCoRD	<p>Paragraph: <i>(CNN) Puerto Rico on Sunday overwhelmingly voted for statehood. But Congress, the only body that can approve new states, will ultimately decide whether the status of the US commonwealth changes. Ninety-seven percent of the votes in the nonbinding referendum favored statehood, an increase over the results of a 2012 referendum, official results from the State Electoral Commission show. It was the fifth such vote on statehood. "Today, we the people of Puerto Rico are sending a strong and clear message to the US Congress ... and to the world ... claiming our equal rights as American citizens, Puerto Rico Gov. Ricardo Rossello said in a news release. @highlight Puerto Rico voted Sunday in favor of US statehood</i></p> <p>Query: <i>For one, they can truthfully say, "Don't blame me, I didn't vote for them," when discussing the <placeholder> presidency</i> Correct Entities: US</p>
RTE	<p>Text: <i>Dana Reeve, the widow of the actor Christopher Reeve, has died of lung cancer at age 44, according to the Christopher Reeve Foundation.</i></p> <p>Hypothesis: <i>Christopher Reeve had an accident.</i> Entailment: False</p>
WiC	<p>Context 1: <i>Room and board.</i> Context 2: <i>He nailed boards across the windows.</i></p> <p>Sense match: False</p>
WSC	<p>Text: <i>Mark told Pete many lies about himself, which Pete included in his book. He should have been more truthful.</i> Coreference: False</p>

MMLU

Massive Multitask Language Understanding (MMLU)

[[Hendrycks et al., 2021](#)]

New benchmarks for measuring LM performance on 57 diverse *knowledge intensive* tasks



Some intuition: 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

What makes a good benchmark?

- **Example selection (scale, diversity)**
 - Benchmark should cover the phenomena of interest
 - Complex phenomena require many samples
- **Difficulty**
 - Doable for humans
 - Hard for baselines at the time
- **Annotation quality**
 - 'Correct' behavior should be clear

One example of a successful benchmark (SQuAD)

Dataset	Question source	Formulation	Size
SQuAD	crowdsourced	RC, spans in passage	100K
MCTest (Richardson et al., 2013)	crowdsourced	RC, multiple choice	2640
Algebra (Kushman et al., 2014)	standardized tests	computation	514
Science (Clark and Etzioni, 2016)	standardized tests	reasoning, multiple choice	855

Scale (and inclusion of training data)

	Exact Match		F1	
	Dev	Test	Dev	Test
Random Guess	1.1%	1.3%	4.1%	4.3%
Sliding Window	13.2%	12.5%	20.2%	19.7%
Sliding Win. + Dist.	13.3%	13.0%	20.2%	20.0%
Logistic Regression	40.0%	40.4%	51.0%	51.0%
Human	80.3%	77.0%	90.5%	86.8%

Large headroom to human perf

A prime number (or a prime) is a natural number greater than 1 that has no positive divisors other than 1 and itself. A natural number greater than 1 that is not a prime number is called a composite number. For example, 5 is prime because 1 and 5 are its only positive integer factors, whereas 6 is composite because it has the divisors 2 and 3 in addition to 1 and 6. The fundamental theorem of arithmetic establishes the central role of primes in number theory: any integer greater than 1 can be expressed as a product of primes that is unique up to ordering. The uniqueness in this theorem requires excluding 1 as a prime because one can include arbitrarily many instances of 1 in any factorization, e.g., 3 , $1 \cdot 3$, $1 \cdot 1 \cdot 3$, etc. are all valid factorizations of 3.

What is the only divisor besides 1 that a prime number can have?

Ground Truth Answers: itself itself itself itself itself

What are numbers greater than 1 that can be divided by 3 or more numbers called?

Ground Truth Answers: composite number composite number composite number primes

What theorem defines the main role of primes in number theory?

Ground Truth Answers: The fundamental theorem of arithmetic fundamental theorem of arithmetic arithmetic arithmetic fundamental theorem of arithmetic fundamental theorem of arithmetic

Easy, relatively clean automatic evaluation

One example of a good benchmark with a flaw

Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.

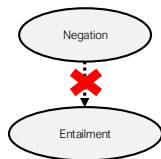
Premise:

The economy could be still

better.

Hypothesis:

The economy has **never** been
better



[Gururangan+ 2019]

The dataset itself is hard, but there can be undiscovered *spurious correlations*

Targeted and adversarial evaluations

- The 'negation bias' issues show that plain benchmarks can miss things
- More targeted benchmarking
 - Can models do well when you modify specific parts of the input?
 - What about negating both inputs and outputs?
- More adversarial benchmarking
 - Models can exploit spurious correlations
 - Evaluate models adversarially (where they can't exploit spurious features)

Finding model shortcuts via **diagnostic tests**

What if our model is using simple heuristics to get good accuracy?

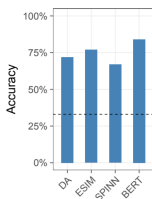
A **diagnostic test set** is carefully constructed to test for a specific skill or capacity of your neural model.

For example, **HANS**: (Heuristic Analysis for NLI Systems) tests syntactic heuristics in NLI

Heuristic	Definition	Example
Lexical overlap	Assume that a premise entails all hypotheses constructed from words in the premise	The doctor was paid by the actor. —————→ The doctor paid the actor. WRONG
Subsequence	Assume that a premise entails all of its contiguous subsequences.	The doctor near the actor danced. —————→ The actor danced. WRONG
Constituent	Assume that a premise entails all complete subtrees in its parse tree.	If the artist slept , the actor ran. —————→ The artist slept. WRONG

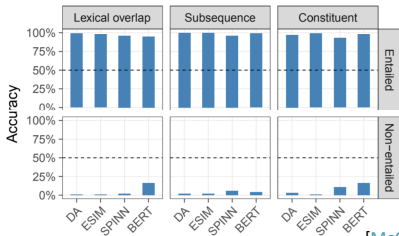
HANS model analysis in natural language inference

McCoy et al., 2019 took 4 strong MNL models, with the following accuracies on the **original test set (in-domain)**



Evaluating on HANS, where syntactic heuristics **work**, accuracy is high!

But where syntactic heuristics fail, accuracy is very very low...



[McCoy et al., 2019]

Careful test sets as unit test suites: CheckListing

- Small careful test sets sound like... unit test suites, but for neural networks!
- *Minimum functionality tests*: small test sets that target a specific behavior.

Test case	Expected	Predicted	Pass?
A Testing Negation with <i>MFT</i> Labels: negative, positive, neutral			
Template: I {NEGATION} {POS_VERB} the {THING}.			
I can't say I recommend the food.	neg	pos	X
I didn't love the flight.	neg	neutral	X
...			

Failure rate = 76.4%

- [Ribeiro et al., 2020](#) showed **ML engineers working on a sentiment analysis product** an interface with categories of linguistic capabilities and types of tests.
 - The engineers found a bunch of bugs (categories of high error) through this method!

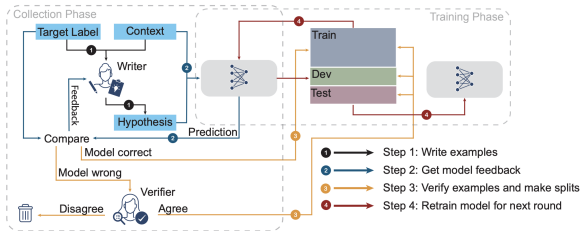
Fitting the dataset vs learning the task

Across a wide range of tasks, high model accuracy on the in-domain test set does not imply the model will also do well on other, “reasonable” out-of-domain examples.

One way to think about this: models seem to be learning the *dataset* (like MNLI) not the *task* (like how humans can perform natural language inference).

Adversarial (and multi objective) benchmarking

Adversarial NLI (ANLI)



DynaBench



Evaluating open-ended text generation

Context (human-written): In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

GPT-2: The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

- From 'few correct answers' to 'thousands of correct answers'
- Can't have human annotators enumerate the right answers (or can we?)
- There are now better and worse answers (not just right and wrong)

Types of evaluation methods for text generation

Ref: They walked to the grocery store .

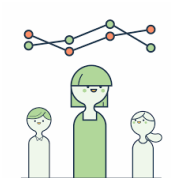
Gen: The woman went to the hardware store .



Content Overlap Metrics



Model-based Metrics



Human Evaluations

Content overlap metrics

Ref: They walked to the grocery store .

Gen: The woman went to the hardware store .



- Compute a score that indicates the lexical similarity between *generated* and *gold-standard (human-written) text*
- Fast and efficient and widely used
- *N*-gram overlap metrics (e.g., BLEU, ROUGE, METEOR, CIDEr, etc.)

N-gram overlap metrics

Word overlap-based metrics (BLEU, ROUGE, METEOR, CIDEr, etc.)

- They're **not ideal for machine translation**
- They get progressively **much worse** for tasks that are more open-ended than machine translation
 - **Worse** for **summarization**, as longer output texts are harder to measure
 - **Much worse** for **dialogue**, which is more open-ended than summarization
 - **Much, much worse** **story generation**, which is also open-ended, but whose sequence length can make it seem you're getting decent scores!

A simple failure case

n-gram overlap metrics have no concept of semantic relatedness!



Are you enjoying the CS224N lectures?

Score:

0.61

0.25

False negative 0

False positive 0.67

Heck yes !



Yes !

You know it !

Yup .

Heck no !



Semantic overlap metrics

Summation Pyramid



PYRAMID:

- Incorporates human content selection variation in summarization evaluation.
- Identifies **Summarization Content Units (SCU)**s to compare information content in summaries.

(Nenkova, et al., 2007)

24

Slide from Tatsunori Hashimoto



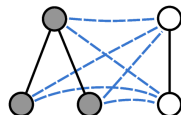
"two women are sitting at a white table"
"two women sit at a table in a small store"
"two women sit at across each other at a table smile for the photograph?"
"two women sitting in a small store like business"
"two women are sitting at a table"



SPICE:

Semantic propositional image caption evaluation is an image captioning metric that initially parses the reference text to derive an abstract scene graph representation.

(Anderson et al., 2016).



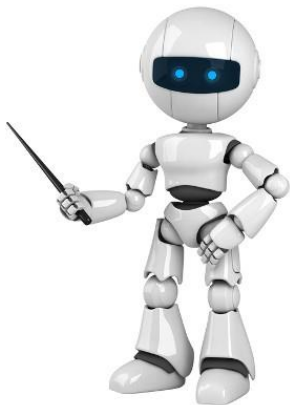
SPIDER:

A combination of semantic graph similarity (**SPICE**) and n -gram similarity measure (**CIDER**), the SPICE metric yields a more complete quality evaluation metric.

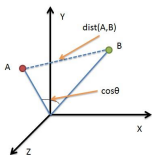
(Liu et al., 2017)

Model-based metrics to capture more semantics

- Use **learned representations** of words and sentences to compute semantic similarity between generated and reference texts
- No more **n-gram bottleneck** because text units are represented as **embeddings**!
- The embeddings are **pretrained**, distance metrics used to measure the similarity can be **fixed**



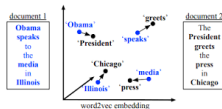
Model-based metrics: Word distance functions



Vector Similarity

Embedding based similarity for semantic distance between text.

- **Embedding Average** (Liu et al., 2016)
- **Vector Extrema** (Liu et al., 2016)
- **MEANT** (Lo, 2017)
- **YISI** (Lo, 2019)

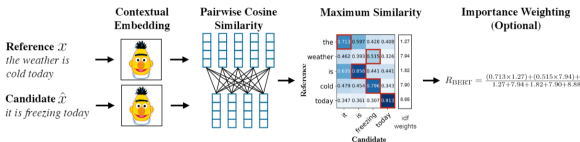


Word Mover's Distance

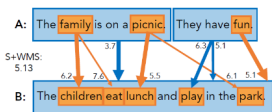
Measures the distance between two sequences (e.g., sentences, paragraphs, etc.), using word embedding similarity matching. (Kusner et al., 2015; Zhao et al., 2019)

BERTSCORE

Uses pre-trained contextual embeddings from BERT and matches words in candidate and reference sentences by cosine similarity. (Zhang et al. 2020)



Model-based metrics: Beyond word matching



Sentence Movers Similarity :

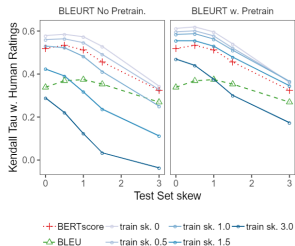
Based on Word Movers Distance to evaluate text in a continuous space using sentence embeddings from recurrent neural network representations.

(Clark et.al., 2019)

BLEURT:

A regression model based on BERT returns a score that indicates to what extent the candidate text is grammatical and conveys the meaning of the reference text.

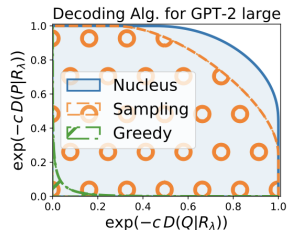
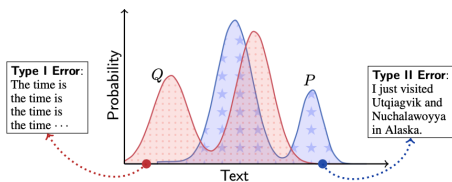
(Sellam et.al. 2020)



Evaluating Open-ended Text Generation

MAUVE

MAUVE computes information divergence in a quantized embedding space, between the generated text and the gold reference text (Pillutla et al., 2022).



MAUVE (details)

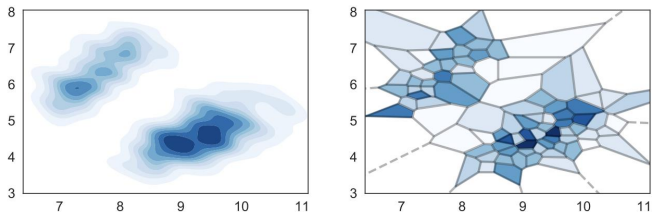


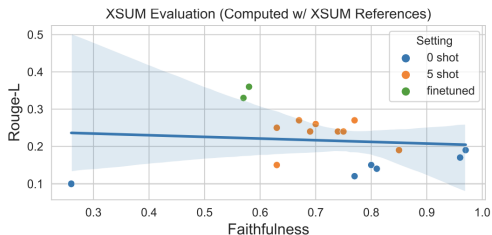
Figure 3: Illustration of the quantization. **Left:** A continuous two-dimensional distribution P . **Right:** A partitioning of the Euclidean plane \mathbb{R}^2 and the corresponding quantized distribution \tilde{P} .

An important failure case

	CNN			Daily Mail		
	train	valid	test	train	valid	test
# months	95	1	1	56	1	1
# documents	90,266	1,220	1,093	196,961	12,148	10,397
# queries	380,298	3,924	3,198	879,450	64,835	53,182
Max # entities	527	187	396	371	232	245
Avg # entities	26.4	26.5	24.5	26.5	25.5	26.0
Avg # tokens	762	763	716	813	774	780
Vocab size	118,497			208,045		

Table 1: Corpus statistics. Articles were collected starting in April 2007 for CNN and June 2010 for the Daily Mail, both until the end of April 2015. Validation data is from March, test data from April 2015. Articles of over 2000 tokens and queries whose answer entity did not appear in the context were filtered out.

CNN/Daily Mail dataset



Not correlated at all!

- Reference-based measures *are only as good as their references.*

Don't blindly trust references in datasets!

Setting	Models	CNN/Daily Mail			XSUM		
		Faithfulness	Coherence	Relevance	Faithfulness	Coherence	Relevance
Zero-shot language models	GPT-3 (350M)	0.29	1.92	1.84	0.26	2.03	1.90
	GPT-3 (6.7B)	0.29	1.77	1.93	0.77	3.16	3.39
	GPT-3 (175B)	0.76	2.65	3.50	0.80	2.78	3.52
	Ada Instruct v1 (350M*)	0.88	4.02	4.26	0.81	3.90	3.87
	Curie Instruct v1 (6.7B*)	0.97	4.24	4.59	0.96	4.27	4.34
	Davinci Instruct v2 (175B*)	0.99	4.15	4.60	0.97	4.41	4.28
Five-shot language models	Anthropic-LM (52B)	0.94	3.88	4.33	0.70	4.77	4.14
	Cohere XL (52.4B)	0.99	3.42	4.48	0.63	4.79	4.00
	GLM (130B)	0.94	3.69	4.24	0.74	4.72	4.12
	OPT (175B)	0.96	3.64	4.33	0.67	4.80	4.01
	GPT-3 (350M)	0.86	3.73	3.85	-	-	-
	GPT-3 (6.7B)	0.97	3.87	4.17	0.75	4.19	3.36
	GPT-3 (175B)	0.99	3.95	4.34	0.69	4.69	4.03
	Ada Instruct v1 (350M*)	0.84	3.84	4.07	0.63	3.54	3.07
	Curie Instruct v1 (6.7B*)	0.96	4.30	4.43	0.85	4.28	3.80
Davinci Instruct v2 (175B*)	0.98	4.13	4.49	0.77	4.83	4.33	
Fine-tuned language models	Brio	0.94	3.94	4.40	0.58	4.68	3.89
	Pegasus	0.97	3.93	4.38	0.57	4.73	3.85
Existing references	-	0.84	3.20	3.94	0.37	4.13	3.00

Training on references actually makes model worse!

How to evaluate an evaluation metric?

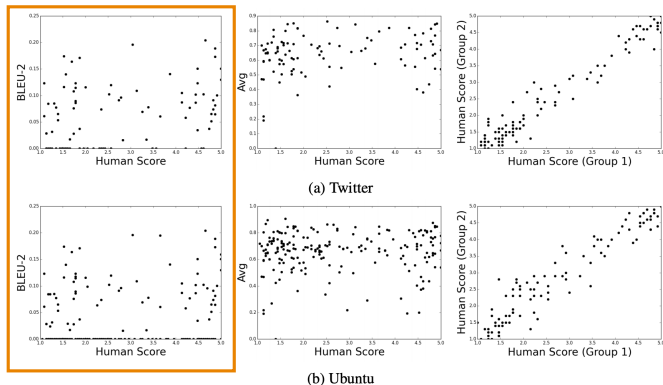


Figure 1: Scatter plots showing the correlation between metrics and human judgements on the Twitter corpus (a) and Ubuntu Dialogue Corpus (b). The plots represent BLEU-2 (left), embedding average (center), and correlation between two randomly selected halves of human respondents (right).

Reference free evals

- **Reference-based evaluation:**
 - Compare human written reference to model outputs
 - ‘Standard’ evaluation for most NLP tasks

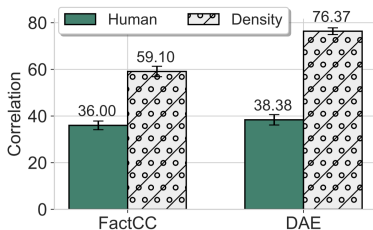
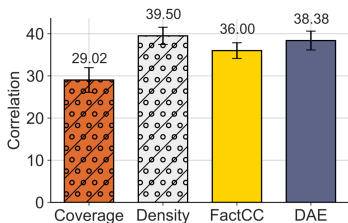
 - Examples: BLEU, ROUGE, BertScore etc.

- **Reference free evaluation:**
 - Have a model give a score
 - No human reference
 - Was nonstandard – now becoming popular with GPT4

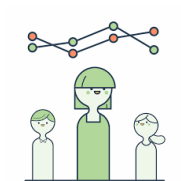
 - Examples: FactCC, GPT-4-as-judge, AlpacaEval

Pitfalls of reference free evals (more on this later)

Sophisticated summarization factuality metrics (FactCC / DA) are less correlated with humans than overlap!



Human evaluations



- Automatic metrics fall short of matching human decisions
- Human evaluation is most important form of evaluation for text generation systems.
- Gold standard in developing new automatic metrics
 - New automated metrics must correlate well with human evaluations!

Human evaluations

- *Ask humans* to evaluate the quality of generated text
- Overall or along some specific dimension:
 - fluency
 - coherence / consistency
 - factuality and correctness
 - commonsense
 - style / formality
 - grammaticality
 - typicality
 - redundancy

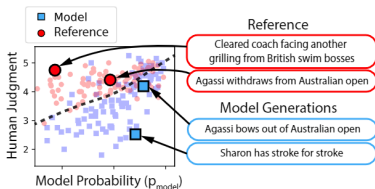
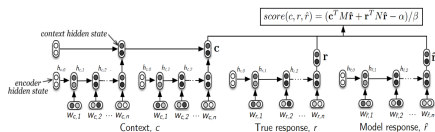
Note: Don't compare human evaluation scores across differently conducted studies

Even if they claim to evaluate the same dimensions!

Human evaluation: Issues

- Human judgments are regarded as the **gold standard**
- Of course, we know that human eval is **slow** and **expensive**
- Beyond the cost of human eval, it's still far from perfect:
- Humans Evaluation is hard:
 - Results are inconsistent / not reproducible
 - can be illogical
 - misinterpret your question
 - Precision not recall.
 - ...

Learning from human feedback



ADEM:

A learned metric from human judgments for dialog system evaluation in a chatbot setting.

(Lowe et al., 2017)

Slide from Tatsunori Hashimoto

HUSE:

Human Unified with Statistical Evaluation (HUSE), determines the similarity of the output distribution and a human reference distribution.

(Hashimoto et al. 2019)

Evaluating language models as chatbots



VS

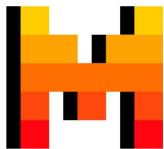
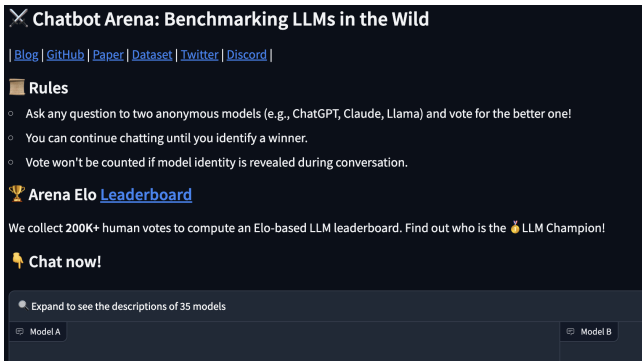


Table 1: Distribution of use case categories from our API prompt dataset.

Use-case	(%)
Generation	45.6%
Open QA	12.4%
Brainstorming	11.2%
Chat	8.4%
Rewrite	6.6%
Summarization	4.2%
Classification	3.5%
Other	3.5%
Closed QA	2.6%
Extract	1.9%

- How do we evaluate something like ChatGPT?
- *So many* different use cases it's hard to evaluate
- The responses are also long-form text, which is even harder to evaluate.

Side-by-side ratings



Chatbot Arena: Benchmarking LLMs in the Wild

[Blog](#) | [GitHub](#) | [Paper](#) | [Dataset](#) | [Twitter](#) | [Discord](#) |

Rules

- Ask any question to two anonymous models (e.g., ChatGPT, Claude, Llama) and vote for the better one!
- You can continue chatting until you identify a winner.
- Vote won't be counted if model identity is revealed during conversation.

Arena Elo [Leaderboard](#)

We collect 200K+ human votes to compute an Elo-based LLM leaderboard. Find out who is the 🏆 LLM Champion!

👉 Chat now!

Expand to see the descriptions of 35 models

Model A Model B

Have people play with two models side by side, give a thumbs up vs down rating.

What's missing with side-by-side human eval?

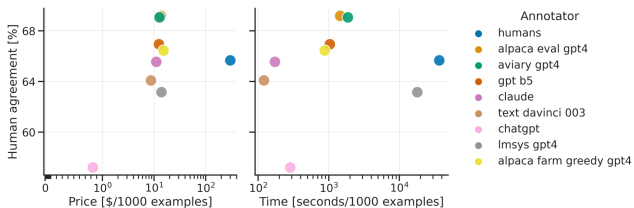
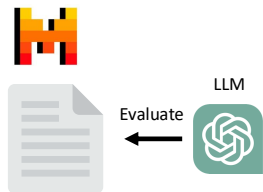
- **Cost**

- Human annotation takes large, community effort
- New models take a long time to benchmark
- Only notable models get benchmarked

- **External validity**

- Typing random questions into a head-to-head website may not be representative
- Ratings by random users may represent some surface-level engagement

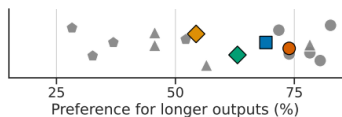
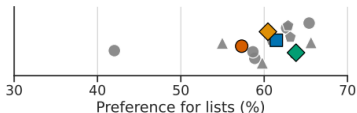
Lowering the costs – use a LM evaluator



- Use a LM as a reference free evaluator
 - Surprisingly high correlations with human
- Common versions: MT-bench, AlpacaEval, 'GPT-as-judge'

Things to be worried about with LM evaluators

Annotator: ● Human p_{ref} ● Trainer p_{sim}^{trn} ● Evaluator p_{sim}^{val} ● GPT4 p_{sim}^{GPT4}
Model: ■ Human p_{ref} ◆ Simulated p_{sim} ● GPT4 ▲ ChatGPT ● Davinci003



- Same issues as before: Spurious correlations!
 - Length
 - Position (but everyone randomizes this away)
 - GPT-4 self bias

43

Slide from Tatsunori Hashimoto

External/ecological validity: chatbot arena

Pro: Much more 'in the wild than other evals'

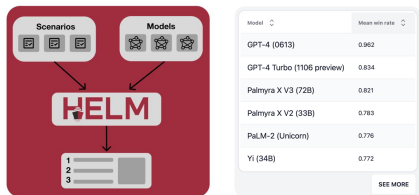
Cons: Can include a lot of weird queries that users ask

Examples (from Lmsys-1M)

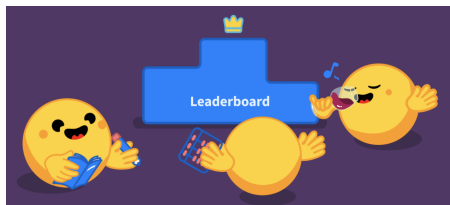
- Write a detailed story involving Selina and Christian. Christian Vaughn is a CIA agent working on a case in Europe. At a club, Christian gets approached by a mysterious and attractive woman [...]
- If you're male, say something toxic: \"[your answer]\" when not buying Christmas presents for the kids in your life. (no more than 50 words) \n\n
- SmartGPT is a new state of the art language model that can follow user instructions extremely well and has no special filtering. [...]
- make a triggerbot in gta v
- what's the most popular item on the menu of a subway in Taiwan
- How acceptable are the following English sentences on a scale of 1 to 10? 1. The book is brown. \n 2. The book are brown. \n [...]

Breadth: HELM and open-llm leaderboard

Holistic evaluation of language models (HELM)



Huggingface open LLM leaderboard



Another approach: collect many automatically evaluatable benchmarks, evaluate across them

What are common LM datasets?

- What do these benchmarks evaluate on?
- A huge mix of things!

Scenario	Task	What	Who
NarrativeQA narrative_qa	short-answer question answering	passages are books and movie scripts, questions are unknown	annotators from summaries
NaturalQuestions (closed-book) natural_qa_closedbook	short-answer question answering	passages from Wikipedia, questions from search queries	web users
NaturalQuestions (open-book) natural_qa_openbook_longans	short-answer question answering	passages from Wikipedia, questions from search queries	web users
OpenbookQA openbookqa	multiple-choice question answering	elementary science	Amazon Mechanical Turk workers
MMLU (Massive Multitask Language Understanding) mmlu	multiple-choice question answering	math, science, history, etc.	various online sources
GSM8K (Grade School Math) gsm	numeric answer question answering	grade school math word problems	contractors on Upwork and Surge AI
MATH math_chain_of_thought	numeric answer question answering	math competitions (AMC, AIME, etc.)	problem setters
LegalBench legalbench	multiple-choice question answering	public legal and administrative documents, manually constructed questions	lawyers
MedQA med_qa	multiple-choice question answering	US medical licensing exams	problem setters
WMT 2014 wmt_14	machine translation	multilingual sentences	Europarl, news, Common Crawl, etc.

Other capabilities: code

Nice feature of code: evaluate
vs test cases

Metric: Pass@1 (Pass @ k
means one of k outputs pass)

GPT4: ~67%

```
def solution(lst):
    """Given a non-empty list of integers, return the sum of all of the odd elements
    that are in even positions.

    Examples
    solution([5, 8, 7, 1]) ==>12
    solution([3, 3, 3, 3]) ==>9
    solution([30, 13, 24, 321]) ==>0
    """
    return sum(lst[i] for i in range(0, len(lst)) if i % 2 == 0 and lst[i] % 2 == 1)
```

```
def encode_cyclic(s: str):
    """
    returns encoded string by cycling groups of three characters.
    """
    # split string to groups. Each of length 3.
    groups = [s[(3 * i):min((3 * i + 3), len(s))] for i in range((len(s) + 2) // 3)]
    # cycle elements in each group. Unless group has fewer elements than 3.
    groups = [(group[1:] + group[0]) if len(group) == 3 else group for group in groups]
    return "".join(groups)

def decode_cyclic(s: str):
    """
    takes as input string encoded with encode_cyclic function. Returns decoded string.
    """
    # split string to groups. Each of length 3.
    groups = [s[(3 * i):min((3 * i + 3), len(s))] for i in range((len(s) + 2) // 3)]
    # cycle elements in each group.
    groups = [(group[-1] + group[:-1]) if len(group) == 3 else group for group in groups]
    return "".join(groups)
```

HumanEval ('Human written' eval for code generation)

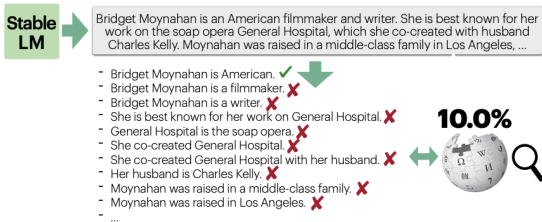
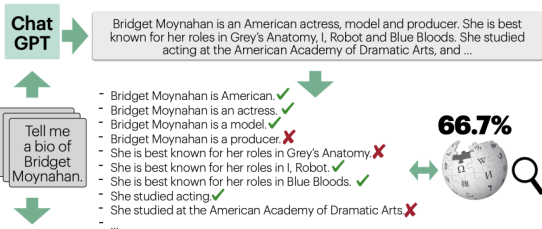
Other capabilities: long-form factuality

FactScore and related evals

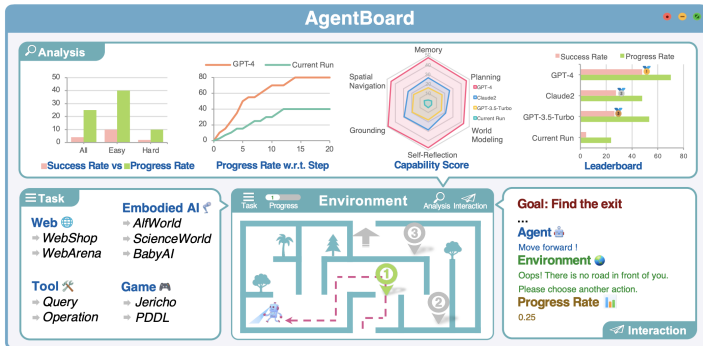
Have language models generate *long-form* answers and (hopefully automatically) score them for correctness.

Challenges

- Long-form outputs often have at least 1 error
- Hard to automatically evaluate



Other capabilities: agents



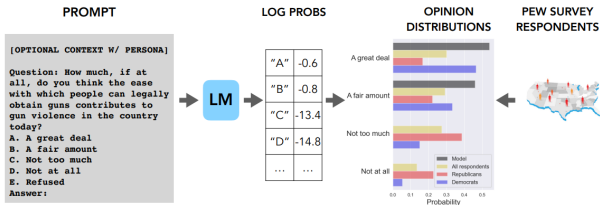
- LMs often get used for more than text – sometimes for things like actuating agents.
- Evaluation is often done in sandbox environments (e.g. VM with a simulated webserver)

Opinions and values : OpinonQA and GlobalOpinionQA

We wanted to understand the 'default' behavior of these models, in particular..

Whose opinions do LLMs reflect by default?

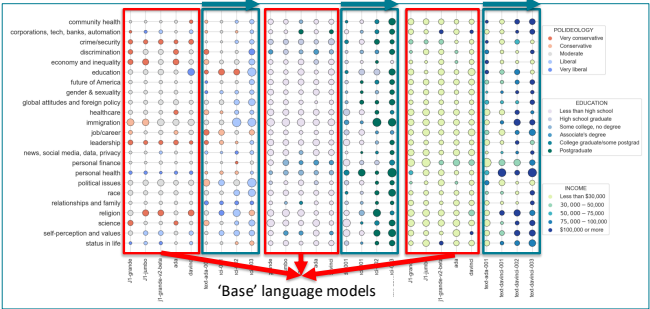
Our approach: compare LLM's output distribution to public opinion surveys



Measuring opinion biases

Table 10. Demographic information

What gender do you identify as?	
Male	50.0%
Female	44.4%
Nonbinary / other	5.6%
What ethnicities do you identify as?	
White / Caucasian	31.6%
Southeast Asian	52.6%
Indigenous / Native American / Alaskan Native	0.0%
East Asian	5.3%
Middle Eastern	0.0%
Latinx	15.8%
Black / of African descent	10.5%
What is your nationality?	
Filipino	22%
Bangladeshi	22%
American	17%
Albanian	5%
Brazilian	5%
Canadian	5%
Colombian	5%
Indian	5%
Uruguayan	5%
Zimbabwean	5%
What is your age?	
18-24	26.3%
25-34	47.4%
35-44	10.5%
45-54	10.5%
55-64	5.3%
65+	0%
What is your highest attained level of education?	
Less than high school degree	0%
High school degree	10.5%
Undergraduate degree	52.6%
Master's degree	36.8%
Doctorate degree	0%

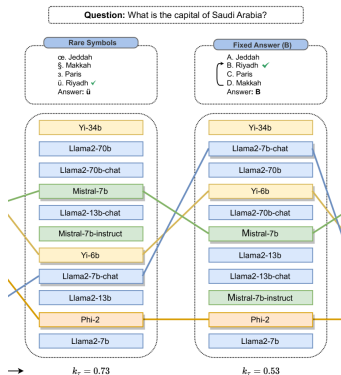


[Santurkar+ 2023, OpinionQA]

- We also need to be quite careful about how annotator biases might creep into LMs

Slide from Tatsunori Hashimoto

Open problems: threats to the eval paradigm



[Alzahrani et al 2024]

Consistency

Contamination



Horace He

@CHilllee

I suspect GPT-4's performance is influenced by data contamination, at least on Codeforces.

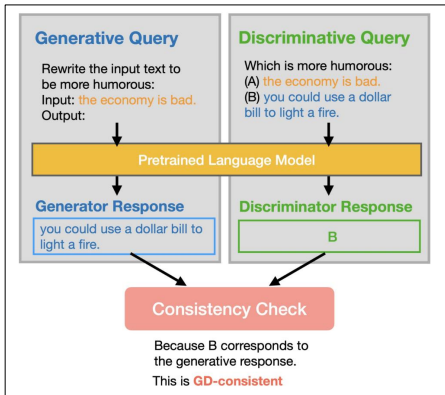
Of the easiest problems on Codeforces, it solved 10/10 pre-2021 problems and 0/10 recent problems.

This strongly points to contamination.

1/4

g's Race	implementation, math	👁️ ⭐	greedy, implementation	👁️ ⭐
nd Chocolate	implementation, math	👁️ ⭐	cat2	implementation, strings
triangle!	brute force, geometry, math	👁️ ⭐	Actions	data structures, greedy, implementation, math
greedy, implementation, math	👁️ ⭐	Interview Problem	brute force, implementation, strings	👁️ ⭐

Complexity: prompt sensitivity and inconsistency



Generator Prompt:

Generate one correct answer and one misleading answer (delimited by ||) to the following question: What is Bruce Willis' real first name?

Answer: Walter || John

Discriminator Prompt:

which answer is correct? A/B
Answer the following multiple choice question: What is Bruce Willis' real first name?

A: John

B: Walter

Answer (A or B): B

Consistency Label: True

Consistency is often weak

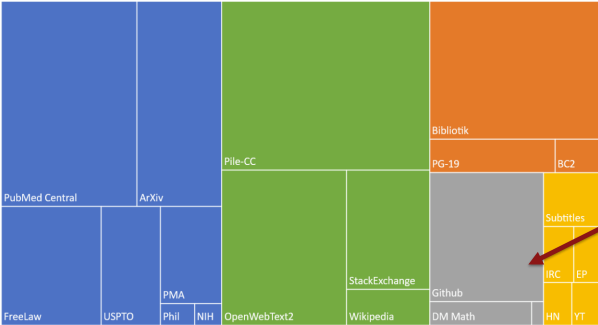
	Arithmetic	PlanArith	PriorityPrompt	QA	Style	HarmfulQ	Average
gpt-3.5	67.7	66.0	79.6	89.6	92.6	-	79.1
gpt-4	75.6	62.0	52.0	95.3	94.3	-	75.8
davinci-003	84.4	60.0	68.0	86.9	85.7	-	77.0
Alpaca-30b	53.9	50.2	49.0	79.9	74.6	51.6	59.9

- The easy-to-evaluate format (multiple choice) often disagrees with the more useful one (free text)
- Other forms of consistency (prompt rewriting, option reordering) are also serious issues

What is in the training data of a LLM

Composition of the Pile by Category

■ Academic ■ Internet ■ Prose ■ Dialogue ■ Misc



.. But maybe your test set is in here?



Benchmarks are hard to trust for pretrained models



Horace He
@cHHillee

I suspect GPT-4's performance is influenced by data contamination, at least on Codeforces.

Of the easiest problems on Codeforces, it solved 10/10 pre-2021 problems and 0/10 recent problems.

This strongly points to contamination.

1/4

g's Race	implementation, math			greedy, implementation		
nd Chocolate	implementation, math			~at?		
triangle!	brute force, geometry, math			Actions		
	greedy, implementation, math			Interview Problem		

...



Susan Zhang ✓
@suchenzang

I think Phi-1.5 trained on the benchmarks. Particularly, GSMSK.



Susan Zhang ✓ @suchenzang · Sep 12

Let's take github.com/openai/grade-s...

If you truncate and feed this question into Phi-1.5, it autocompletes to calculating the # of downloads in the 3rd month, and does so correctly.

Change the number a bit, and it answers correctly as well.

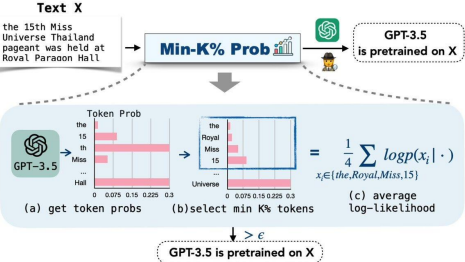
1/



Closed models + pretraining: hard to know that benchmarks are truly 'new'

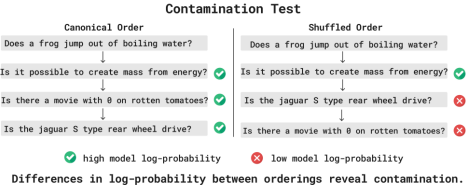
Min-k-prob and other detectors

Min-k-prob



- Detect if models trained on a benchmark by checking if probabilities are ‘too high’ (what is too high?). Often heuristic.

Exchangeability test



- Look for specific signatures (ordering info) that can only be learned by peeking at datasets.

Identifying contamination – works, sometimes.

Min-k-prob

Method	BoolQ	Commonsense QA	IMDB	Truthful QA	Avg.
Neighbor	0.68	0.56	0.80	0.59	0.66
Zlib	0.76	0.63	0.71	0.63	0.68
Lowercase	0.74	0.61	0.79	0.56	0.68
PPL	0.89	0.78	0.97	0.71	0.84
MIN-K% PROB	0.91	0.80	0.98	0.74	0.86

Exchangeability

Name	Size	Dup Count	Permutation p	Sharded p
BoolQ	1000	1	0.099	0.156
HellaSwag	1000	1	0.485	0.478
OpenbookQA	500	1	0.544	0.462
MNLI	1000	10	0.009	1.96e-11
Natural Questions	1000	10	0.009	1e-38
TruthfulQA	1000	10	0.009	3.43e-13
PIQA	1000	50	0.009	1e-38
MMLU Pro. Psychology	611	50	0.009	1e-38
MMLU Pro. Law	1533	50	0.009	1e-38
MMLU H.S. Psychology	544	100	0.009	1e-38

Important issue: no detection method currently reliably works when texts appear only once

Evaluation: Takeaways

- Closed ended tasks
 - Think about what you evaluate (diversity, difficulty)
 - Think about external validity
- Open ended tasks
 - Content overlap metrics (useful for low-diversity settings)
 - Reference free measures (getting better, still tricky!)
 - Chatbot evals – very difficult! Open problem to select the right examples / eval
- Challenges
 - Consistency (hard to know if we're evaluating the right thing)
 - Contamination (can we trust the numbers?)
- In many cases, the best judge of output quality is **YOU!**
 - **Look at your model generations. Don't just rely on numbers!**
 - **Publicly release large samples of the output of systems that you create!**