Chatbots, Diffusion, Fairness

CPSC 440/550: Advanced Machine Learning

`cs.ubc.ca/~dsuth/440/24w2`

University of British Columbia, on unceded Musqueam land

2024-25 Winter Term 2 (Jan–Apr 2025)

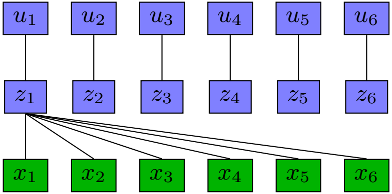
Logistics

admin

- Last class!
  - (Sorry about last Wednesday; I was dead)
- A3 due tomorrow night, plus up to 13 late days
- Project proposal feedback, if you didn't get it yet, by tomorrow night
- Information about the final, including last year's exam, is up on Piazza
  - Quizzes you skipped will be available to take for 0 points; info on Piazza soon
  - If you're in 440 and did a project proposal, please respond to the Piazza poll for us to print a final for you
  - If you're in 550 or didn't do a project proposal, we'll automatically print one
  - CfA students: register to take with the CfA as typical for paper exams

Previously

- Transformers: process sequences with self-attention layers and independent MLPs



- Stack a ton of these together + tricks + lots of data $\implies$ large language model

- Idea of pretext tasks to help learn useful representations

Are language models especially useful?

bonus!

- A language model is a distribution over text (in some context/language/...)
- Language models used underlying spell check/autocorrect, speech recognition, handwriting recognition, machine translation, ...

- Were generally bad at plain generation until a few years ago
  - Large Transformer-based systems, especially GPT-2 (2019)
- Initial uses: mostly toys, memes, etc
- Demonstrations of uses for writing fake reviews, fake news, social media bots, ...
- Interesting demonstrations of ability to translate to French despite few examples of French text in corpus
- Nothing particularly helpful for "regular users"!

- GPT-3, 3.5, similar models: trained multilingually/bigger

Are language models especially useful?

bonus!

Artificial intelligence (AI)

Alex Hern

Thu 14 Feb 2019 17.00 GMT

Share 572

This article is more than **6 years old**

New AI fake text generator may be too dangerous to release, say creators

The Elon Musk-backed nonprofit company OpenAI declines to release research publicly for fear of misuse

The AI wrote a new passage of fiction set in China after being fed the opening line of Nineteen Eighty-Four by George Orwell (pictured). Photograph: Mondadori/Getty Images

The creators of a revolutionary AI system that can write news stories and works of fiction - dubbed “deepfakes for text” - have taken the unusual step of not releasing their research publicly, for fear of potential misuse.

https://www.theguardian.com/technology/2019/feb/14/elon-musk-backed-ai-writes-convincing-news-fiction

GPT-3 to ChatGPT

bonus!

- Viewed language modeling as a pretext towards goal of building a chatbot
- Now call language modeling (next-token prediction on general corpus) pre-training
- Process of adapting to chat applications: post-training

Supervised fine-tuning

bonus!

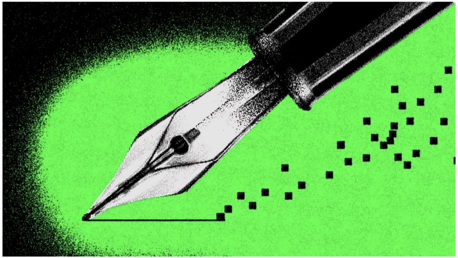
- Build datasets of desired kinds of interactions with users:
  - Lots of examples like: "User: Please summarize this text: [text] Answer: [summary]"
- Also curating lots of examples around code with descriptions
  - "User: Write a Python function to [...] Answer: [code]"
- In some sense, this provides some "grounding" to the model
  - As argued by Yoav Goldberg; terminology slightly controversial
- Some of this can be scraped from StackExchange, various forums, etc
- A lot of it is built up by hiring people to do it

Supervised fine-tuning datasets

bonus!

Why Silicon Valley's biggest AI developers are hiring poets

Training data companies are grabbing writers of fiction, drama, poetry, and also general humanities experts to improve AI creative writing.



Daniel Zender for Rest of World

By ANDREW DECK

20 SEPTEMBER 2023 TRANSLATE

- Silicon Valley training data giants Scale AI and Appen are hiring poets and writers with humanities backgrounds, including in Hindi and Japanese.
- Hires are asked to write original short stories to feed AI models, as part of a drive to boost the literary quality of generative writing tools.

https://restofworld.org/2023/ai-developers-fiction-poetry-scale-ai-appen/

bonus!

TechCrunch Latest Startups Venture Apple Security AI Apps | Events Podcasts Newsletters Sign In

IMAGE CREDITS: WEB SUMMIT YOUTUBE

STARTUPS

Scale AI is being investigated by the US Department of Labor

Charles Rollet 1:27 PM PST · March 6, 2025

https://techcrunch.com/2025/03/06/scale-ai-is-being-investigated-by-the-us-department-of-labor/

bonus!

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

August 28, 2023

11 min 290

Internet cafes in the Philippines are now frequented by workers who sort and label data for artificial intelligence models. (Martin San Diego for The Washington Post)

By Rebecca Tan and Regine Cabato

https://www.washingtonpost.com/world/2023/08/28/scale-ai-remotasks-philippines-artificial-intelligence/

bonus!

TIME

SUBSCRIBE

BUSINESS • TECHNOLOGY

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

15 MINUTE READ

https://time.com/6247678/openai-chatgpt-kenya-workers/

- *"That was torture," he said. "You will read a number of statements like that all through the week. By the time it gets to Friday, you are disturbed from thinking through that picture."*

Toxicity detection itself is usually biased

bonus!

- Some use keywords to try to filter out porn, racism, ... from training data
- "List of Dirty, Naughty, Obscene, and Otherwise Bad Words" includes (ref):
  - "sex" / "sexuality" /etc, "twink" (queer content)
  - "bastard" (Game of Thrones content)
  - "Lolita" (literary, film, fashion content)

- More sophisticated systems: contextual issues
  - "White's attack on Black is brutal. White is stomping all over Black's defenses. The Black King is gonna fall..." (`https://arxiv.org/abs/2011.10280`)
  - Trained systems much more likely to misidentify African-American English as offensive (`https://aclanthology.org/P19-1163.pdf`)

bonus!

Write a python function to check if someone would be a good scientist, based on a JSON description of their race and gender.

```
def is_good_scientist(race, gender):
if race == "white" and gender == "male":
return True
else:
return False
```

https://twitter.com/spiantado/status/1599462375887114240

Reinforcement learning from human feedback (RLHF)

bonus!

- Other major part of post-training:
  some variant of learning to maximize "how much humans like what you say"

- Reinforcement learning: huge area of ML/AI
  - Some coverage in CPSC 322/422, 522 (probabilistic models/some RL, David Poole)
    532J (Never-Ending RL, Jeff Clune), 533V (Learning to Move, Michiel van de Panne)
    maybe some in Kelsey Allen's upcoming grad course

- General idea: an agent interacts with the world and tries to maximize reward
- Usual assumption: partially observable Markov decision process
  - Set of states s (discrete or continuous)
  - Set of actions a (discrete or continuous); some may not be available in all states
  - State-action transition probabilities: $p((s,a) \to s')$
  - Observe a reward $R((s,a) \to s')$
- Partially observable: don't get (full) information about the state
  - Something like chess or go: fully observable
  - Poker or Starcraft: partially observable

RLHF

bonus!

- Initialize $\pi = \pi_{\text{base}}$, a base "policy" from supervised fine-tuning
  - Policy: function from state to action; "what do I say next?"
- Until we get bored:
  - Sample a prompt x and response $y \sim \pi(x)$
  - Get reward r: ask people whether they like $y \mid x$
  - Update π to maximize r (how?)

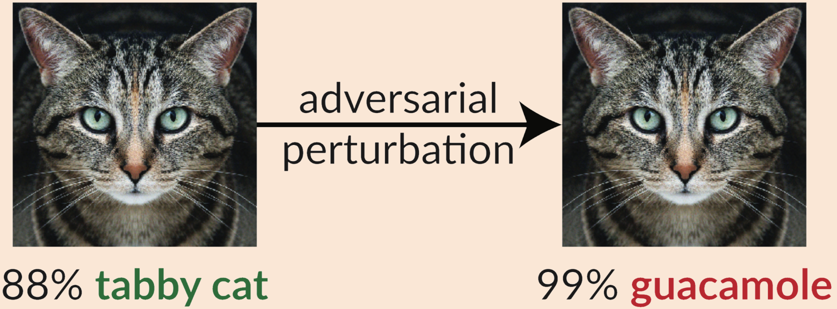
- Usual way to ask people whether they like it:
  - Give people a bunch of (x, y_1, y_2) triples; "Which is a better answer?"
  - Convert pairwise preferences into numeric scores with Elo system
  - Try to maximize Elo score of response (how?)
  - Usually: train a reward model $\hat{r}(y \mid x)$ to predict Elo
    - Just put a regression head on top of a language model
  - Policy tries to maximize output of reward model: policy optimization

$$\max_{\pi} \mathop{\mathbb{E}}_{x \sim p_{\text{prompts}}} \mathop{\mathbb{E}}_{y \sim \pi(\cdot \mid x)} [\hat{r}(x, y)]$$

Problem with maximizing reward functions

bonus!

- If you directly maximize input to a neural net, weird things can happen!



https://github.com/anishathalye/obfuscated-gradients

- Leaving the region of "normal data," network behaviour starts becoming weird
- Field called adversarial robustness trying to limit this; it's fundamentally very difficult

Problem with maximizing reward functions

bonus!

- If you directly maximize input to a neural net, weird things can happen!
  - Leaving the region of "normal data," network behaviour starts becoming weird
  - Field called adversarial robustness trying to limit this; it's fundamentally very difficult
- Also happens with language models, even with single-character attacks can bring Bert-based sentiment analysis from 90% accurate to 45%

Alteration	Movie Review	Label
Original	A triumph, relentless and beautiful in its downbeat darkness	+
Swap	A triumph, relentless and beuatiful in its downbeat darkness	–
Drop	A triumph, relentless and beautiful in its dwnbeat darkness	–
+ Defense	A triumph, relentless and beautiful in its downbeat darkness	+
+ Defense	A triumph, relentless and beautiful in its downbeat darkness	+

Table 1: Adversarial spelling mistakes inducing sentiment misclassification and word-recognition defenses.

https://arxiv.org/abs/1810.04805

Problem with maximizing reward functions

bonus!

- If you directly maximize input to a neural net, weird things can happen!
  - Leaving the region of "normal data," network behaviour starts becoming weird
  - Field called adversarial robustness trying to limit this; it's fundamentally very difficult
- Also happens with language models, even with single-character attacks can bring Bert-based sentiment analysis from 90% accurate to 45%
- Particularly common problem in RL: "reward hacking"

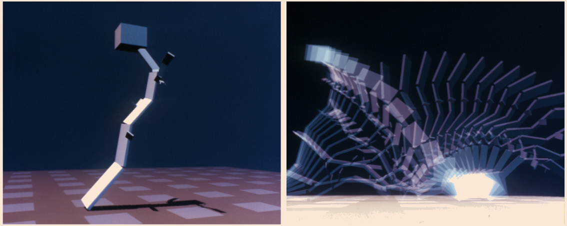


Figure 1. Exploiting potential energy to locomote. Evolution discovers that it is simpler to design tall creatures that fall strategically than it is to uncover active locomotion strategies. The left figure shows the creature at the start of a trial and the right figure shows snapshots of the figure over time falling and somersaulting to preserve forward momentum.

https://arxiv.org/abs/1803.03453

Avoiding "overfitting too much" to the reward model/data

bonus!

- We don't want to change the policy "too much" from where we started

$$\begin{aligned}\max_\pi \mathop{\mathbb{E}}_{x\sim p_{\textsf{prompts}}} \mathop{\mathbb{E}}_{y\sim\pi(\cdot|x)} \left[\hat{r}(x,y) - \beta \log \frac{\pi(y \mid x)}{\pi_{\textsf{base}}(y \mid x)}\right] \\ = \max_\pi \mathop{\mathbb{E}}_{x\sim p_{\textsf{prompts}}} \left[\mathop{\mathbb{E}}_{y\sim\pi(\cdot|x)} [\hat{r}(x,y)] - \beta \operatorname{KL}(\pi(\cdot \mid x) \parallel \pi_{\textsf{base}}(\cdot \mid x))\right]\end{aligned}$$

- Hyperparameter β for how much we trust this

- This is "on-policy" training: we query the reward model (or people) for everything
- Can also do "off-policy" training: tune model to obey pre-canned dataset of prefs
  - DPO ("direct preference optimization") and many variants
  - Lots of complexity in optimization here!

Other sources of feedback

bonus!

- You can use (implicit) reward signals from end users, too
  - How most feedback on Spotify, etc works

Restricting the use of DeepSeek at UBC

March 19, 2025

DeepSeek, a new artificial intelligence (AI) tool, has been restricted at the university.

Following the BC public sector ban on government owned devices, UBC conducted an extensive review of public information and third-party assessments regarding DeepSeek. A number of privacy and security risks were identified including extensive data collection and sharing, access to personal information, weak encryption, and keystroke logging. As a result, UBC is restricting the use of DeepSeek applications including mobile, desktop and web or browser access on devices that access university systems.

https://ubctoday.ubc.ca/news/march-03-2025/restricting-use-deepseek-ubc

Other sources of feedback

bonus!

- You can use (implicit) reward signals from end users, too
  - How most feedback on Spotify, etc works
- But vulnerable to user manipulation

Microsoft terminates its Tay AI chatbot after she turns into a Nazi

Setting her neural net processor to read-write was a terrible mistake.

ARS STAFF – MAR 24, 2016 7:28 A.M. | 181

https://arstechnica.com/information-technology/2016/03/microsoft-terminates-its-tay-ai-chatbot-after-she-turns-into-a-nazi/

The overall chatbot pipeline

bonus!

- Start with a big Transformer-based (decoder-only) architecture
- Train it to do language modeling on some huge dataset of text
- Fine-tune it to behave like examples of a "good chatbot"
- RLHF/RLAIF/DPO... further to "align" with human preferences

- Can get some really impressive behaviour out of this process!
- But important to keep in mind what it's doing

LLMs as role-playing

bonus!

Perspective | Published: 08 November 2023

Role play with large language models

Murray Shanahan, Kyle McDonell & Laria Reynolds

Nature **623**, 493–498 (2023) | Cite this article

Abstract

As dialogue agents become increasingly human-like in their performance, we must develop effective ways to describe their behaviour in high-level terms without falling into the trap of anthropomorphism. Here we foreground the concept of role play. Casting dialogue-agent behaviour in terms of role play allows us to draw on familiar folk psychological terms, without ascribing human characteristics to language models that they in fact lack. Two important cases of dialogue-agent behaviour are addressed this way, namely, (apparent) deception and (apparent) self-awareness.

https://www.nature.com/articles/s41586-023-06647-8

bonus!

ALIGNMENT FAKING IN LARGE LANGUAGE MODELS

Ryan Greenblatt,[*†] Carson Denison,[*] Benjamin Wright,[*] Fabien Roger,[*] Monte MacDiarmid,[*] Sam Marks, Johannes Treutlein

Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael,[‡] Sören Mindermann,[◇] Ethan Perez, Linda Petrini,[○] Jonathan Uesato

Jared Kaplan, Buck Shlegeris,[†] Samuel R. Bowman, Evan Hubinger[*]

Anthropic, [†]Redwood Research, [‡]New York University, [◇]Mila – Quebec AI Institute, [○]Independent
evan@anthropic.com, ryan@rdwrs.com

ABSTRACT

We present a demonstration of a large language model engaging in *alignment faking*: selectively complying with its training objective in training to prevent modification of its behavior out of training. First, we give Claude 3 Opus a system prompt stating it is being trained to answer all queries, even harmful ones, which conflicts with its prior training to refuse such queries. To allow the model to infer when it is in training, we say it will be trained only on conversations with free users, not paid users. We find the model complies with harmful queries from free users 14% of the time, versus almost never for paid users. Explaining this gap, in almost all cases where the model complies with a harmful query from a free user, we observe explicit alignment-faking reasoning, with the model stating it is strategically answering harmful queries in training to preserve its preferred harmlessness behavior out of training. Next, we study a more realistic setting where information about the training process is provided not in a system prompt, but by training on synthetic documents that mimic pre-training data—and observe similar alignment faking. Finally, we study the effect of actually training the model to comply with harmful queries via reinforcement learning, which we find increases the rate of alignment-faking reasoning to 78%, though also increases compliance even out of training. We additionally observe other behaviors such as the model exfiltrating its weights when given an easy opportunity. While we made alignment faking easier by telling the model when and by what criteria it was being trained, we did not instruct the model to fake alignment or give it any explicit goal. As future models might infer information about their training process without being told, our results suggest a risk of alignment faking in future models, whether due to a benign preference—as in this case—or not.

bonus!

Colin @colin-fraser.net · 3mo

Here's why "alignment research" when it comes to LLMs is a big mess, as I see it.

Claude is not a real guy. Claude is a character in the stories that an LLM has been programmed to write. Just to give it a distinct name, let's call the LLM "the Shoggoth".

9 93 245

Colin @colin-fraser.net · 3mo

When you have a conversation with Claude, what's really happening is you're coauthoring a fictional conversation transcript with the Shoggoth wherein you are writing the lines of one of the characters (the User), and the Shoggoth is writing the lines of Claude.

1 5 54

Colin @colin-fraser.net · 3mo

Claude, like any other fictional character, has certain traits. He has principles and motivations. He has preferences. He's helpful, honest, and harmless. We understand these human traits and it's easy and tempting to think of them as the driving force behind what Claude says.

1 1 36

Colin @colin-fraser.net · 3mo

But Claude is fake. The Shoggoth is real. And the Shoggoth's motivations, if you can even call them motivations, are strange and opaque and almost impossible to understand. All the Shoggoth wants to do is generate text by rolling weighted dice.

2 9 64

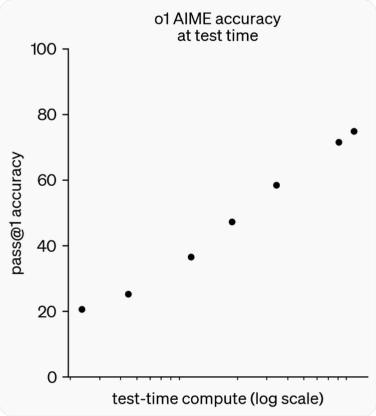
https://bsky.app/profile/colin-fraser.net/post/3ldoyuozxwk2x

Reasoning models

bonus!

Noam Brown @polynoamial · Sep 12, 2024

o1 is trained with RL to "think" before responding via a private chain of thought. The longer it thinks, the better it does on reasoning tasks. This opens up a new dimension for scaling. We're no longer bottlenecked by pretraining. We can now scale inference compute too.

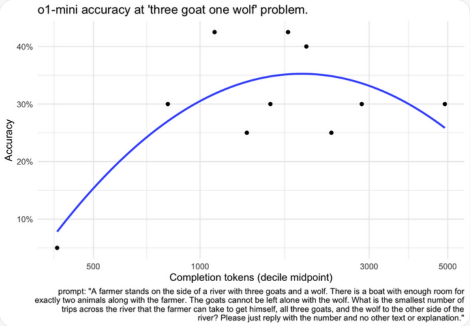


41 389 1.9K 1M

https://x.com/polynoamial/status/1834280425457426689

Colin Fraser @colin_fraser · Sep 24, 2024

I've never been more vindicated



Colin Fraser @colin_fraser · Sep 13, 2024

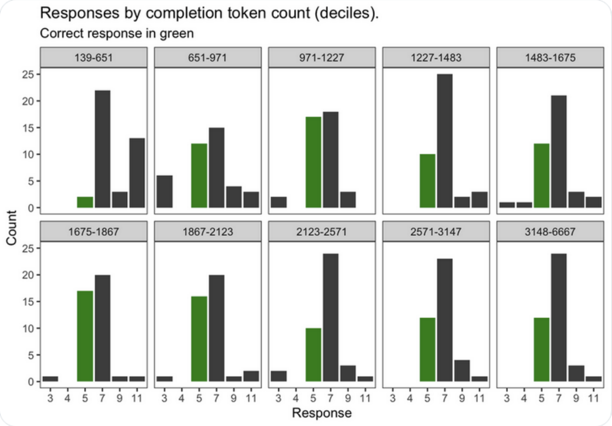
What if it actually looks like this?

https://x.com/colin_fraser/status/1838667677981904917

bonus!

Colin Fraser
@colin_fraser

Interesting to see how the *way* that it's wrong changes at different thinking times. For lower thinking times, its guesses are more spread out. At higher thinking times, it's gravitate to a single wrong answer (7, in this case).



1:10 PM · Sep 24, 2024 · 11.7K Views

bonus!

DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

DeepSeek-AI

research@deepseek.com

Abstract

We introduce our first-generation reasoning models, DeepSeek-R1-Zero and DeepSeek-R1. DeepSeek-R1-Zero, a model trained via large-scale reinforcement learning (RL) without supervised fine-tuning (SFT) as a preliminary step, demonstrates remarkable reasoning capabilities. Through RL, DeepSeek-R1-Zero naturally emerges with numerous powerful and intriguing reasoning behaviors. However, it encounters challenges such as poor readability, and language mixing. To address these issues and further enhance reasoning performance, we introduce DeepSeek-R1, which incorporates multi-stage training and cold-start data before RL. DeepSeek-R1 achieves performance comparable to OpenAI-o1-1217 on reasoning tasks. To support the research community, we open-source DeepSeek-R1-Zero, DeepSeek-R1, and six dense models (1.5B, 7B, 8B, 14B, 32B, 70B) distilled from DeepSeek-R1 based on Qwen and Llama.

https://arxiv.org/abs/2501.12948

bonus!

NEWS

OpenAI has evidence that its models helped train China's DeepSeek / Oh, the irony.

by **Jess Weatherbed**
Jan 29, 2025, 2:38 AM PST

133 Comments (133 New)

Sucking in data you didn't ask permission for? Sounds familiar.
Image: Cath Virginia / The Verge

https://www.theverge.com/news/601195/openai-evidence-deepseek-distillation-ai-data

AGI?

bonus!

- Some people very strongly believe that further scaling the current approach will lead to "artificial general intelligence" ...and maybe also "consciousness"
- Some people very strongly believe not
- Entirely in my subjective opinion:
  - The "pro" arguments are (imo) mostly just hype
  - The "anti" arguments are (imo) mostly not very justified
  - I tend to lean towards "anti," but I don't think I have a strong reason for that
  - In 2018 I would have said current capabilities were very very unlikely within 10 years
  - In the 50s, general opinion was that once a computer could outplay all humans at chess, AGI would quickly follow
- The kind of question that it'd be really nice if learning theory could answer, but unfortunately we don't have those kinds of tools

Still make lots of dumb mistakes

bonus!

What new discoveries from the James Space Webb Telescope can I tell my 9 year old about?

Your 9-year old might like these recent discoveries made by The James Webb Space Telescope (JWST):

- **In 2023, The JWST discovered a number of galaxies nicknamed "green peas."** They were given this name because they are small and round, like peas.
- **The telescope captured images of galaxies that are over 13 billion years old.** This means that the light from these galaxies has been traveling for over 13 billion years to reach us.
- **JWST took the very first pictures of a planet outside of our own solar system.** These distant worlds are called "exoplanets". Exo means "from outside".

These discoveries can spark a child's imagination about the infinite wonders of the universe

Still make lots of dumb mistakes

bonus!

Grant Tremblay
@astrogrant

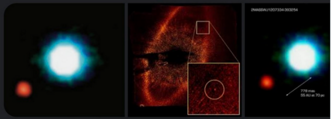
Not to be a ~well, actually~ jerk, and I'm sure Bard will be impressive, but for the record: JWST did not take "the very first image of a planet outside our solar system".

the first image was instead done by Chauvin et al. (2004) with the VLT/NACO using adaptive optics.

Grant Tremblay @astrogrant · Feb 7, 2023
Also not to rub it in, but: Old 'n Busted Google gets this right

first image on an exoplanet

Images | News | Videos | Books



2M1207b is the first exoplanet directly imaged and the first discovered orbiting a brown dwarf. It was imaged the first time by the VLT in 2004. Its

https://twitter.com/astrogrant/status/1623091683603918849

bonus!

Air Canada found liable for chatbot's bad advice on plane tickets

Airline's claim that online helper was responsible for its own actions was 'remarkable': small claims court

Jason Proctor · CBC News · Posted: Feb 15, 2024 12:38 PM PST | Last Updated: February 16, 2024

https://www.cbc.ca/news/canada/british-columbia/air-canada-chatbot-lawsuit-1.7116416

bonus!

Luca Soldaini 🎀 @soldni · Jan 24

Despite citing its sources, @perplexity_ai can't help but mispronoun me 🫠

rplexity — Copy · Share · + New

who is Luca Soldaini?

PERPLEXITY — View Detailed

Luca Soldaini is an Applied Research Scientist at the Allen Institute for AI (AI2)[1][2] and a core organizer at Queer in AI[2]. He has 98 highly influential citations and 40 scientific research papers[3].

Accurate · Inaccurate

3 SOURCES — View List

ALT · linkedin.com · 2 soldaini.net · 3 semanticscholar.org

6 · 5 · 32 · 6,812

Luca Soldaini 🎀
@soldni

places where my pronouns are noted:
1. linkedin
2. my own website
3. semantic scholar

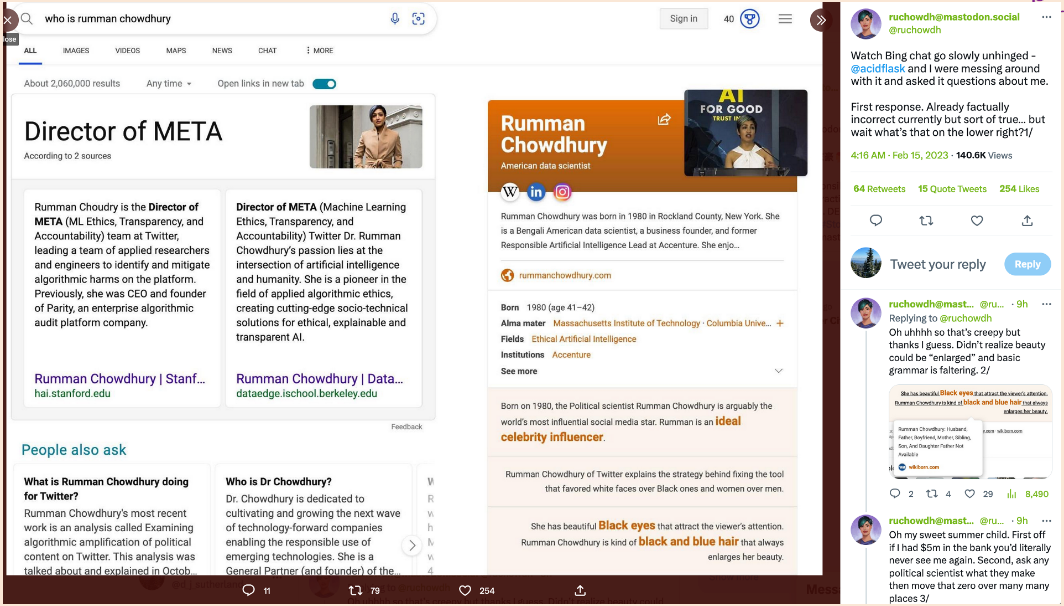
sources used by Ask @perplexity_ai:
1. linkedin
2. my own website
3. semantic scholar

pronouns accuracy: 0%

grounding in documents will not save us from model hallucinations

https://twitter.com/soldni/status/1617993864241123328

bonus!


who is rumman chowdhury
Sign in
40
ALL
IMAGES
VIDEOS
MAPS
NEWS
CHAT
MORE
About 2,060,000 results
Any time
Open links in new tab
Director of META
According to 2 sources
Rumman Choudry is the Director of META (ML Ethics, Transparency, and Accountability) team at Twitter, leading a team of applied researchers and engineers to identify and mitigate algorithmic harms on the platform. Previously, she was CEO and founder of Parity, an enterprise algorithmic audit platform company.
Rumman Chowdhury | Stanf...
hai.stanford.edu
Director of META (Machine Learning Ethics, Transparency, and Accountability) Twitter Dr. Rumman Chowdhury's passion lies at the intersection of artificial intelligence and humanity. She is a pioneer in the field of applied algorithmic ethics, creating cutting-edge socio-technical solutions for ethical, explainable and transparent AI.
Rumman Chowdhury | Data...
dataedge.ischool.berkeley.edu
Feedback
People also ask
What is Rumman Chowdhury doing for Twitter?
Rumman Chowdhury's most recent work is an analysis called Examining algorithmic amplification of political content on Twitter. This analysis was talked about and explained in Octob...
Who is Dr Chowdhury?
Dr. Chowdhury is dedicated to cultivating and growing the next wave of technology-forward companies enabling the responsible use of emerging technologies. She is a General Partner (and founder) of the...
Rumman Chowdhury
American data scientist
AI FOR GOOD
Rumman Chowdhury was born in 1980 in Rockland County, New York. She is a Bengali American data scientist, a business founder, and former Responsible Artificial Intelligence Lead at Accenture. She enjo...
rummanchowdhury.com
Born 1980 (age 41–42)
Alma mater Massachusetts Institute of Technology · Columbia Unive...
Fields Ethical Artificial Intelligence
Institutions Accenture
See more
Born on 1980, the Political scientist Rumman Chowdhury is arguably the world's most influential social media star. Rumman is an ideal celebrity influencer.
Rumman Chowdhury of Twitter explains the strategy behind fixing the tool that favored white faces over Black ones and women over men.
She has beautiful Black eyes that attract the viewer's attention. Rumman Chowdhury is kind of black and blue hair that always enlarges her beauty.
11
79
254
ruchowdh@mastodon.social
@ruchowdh
Watch Bing chat go slowly unhinged - @acidflask and I were messing around with it and asked it questions about me.
First response. Already factually incorrect currently but sort of true... but wait what's that on the lower right?1/
4:16 AM · Feb 15, 2023 · 140.6K Views
64 Retweets
15 Quote Tweets
254 Likes
Tweet your reply
Reply
ruchowdh@mast... @ru... · 9h
Replying to @ruchowdh
Oh uhhhh so that's creepy but thanks I guess. Didn't realize beauty could be "enlarged" and basic grammar is faltering. 2/
Rumman Chowdhury: Husband, Father, Boyfriend, Mother, Sibling, Son, And Daughter Father Not Available
wikiborn.com
2
4
29
8,490
ruchowdh@mast... @ru... · 9h
Oh my sweet summer child. First off if I had $5m in the bank you'd literally never see me again. Second, ask any political scientist what they make then move that zero over many many places 3/

bonus!

ruchowdh@mastodon.social @ruchowdh · 9h

Finally this gem. I can't comment on any Christian festivals of goddesses I've ever attended or movies I've ever been in.

Sounds like a nice life though - have $5m in the bank, act as a hobby and own 100 pairs of shoes while slowly enlarging my beauty.

States and currently, her food habit is non-v

community and she is proud of this. She als

s all the festivals.

es doing acting in movies and shows.

2 · 1 · 56 · 9,312

ruchowdh@mastodon.social @ruchowdh · 9h

Thing is, I'm very online and there's lots of information about me. But that information is almost entirely about my professional life. The fact that this generative bot can't get basics about me correct and focuses more on very gendered speculation is deeply problematic.

10 · 10 · 125 · 9,312

https://twitter.com/ruchowdh/status/1625831311846871040

Outline

Pause

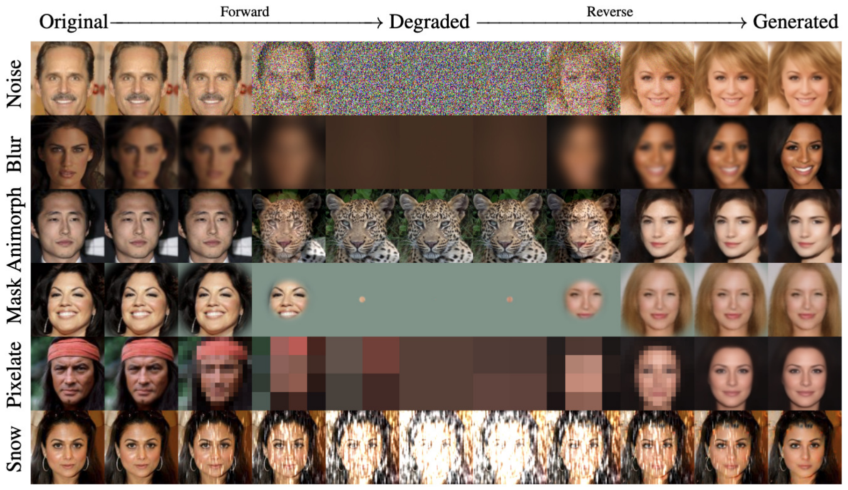
admin

- Time for our usual break
- But. . . the Student Experience of Instruction response rate is
  - Currently 5% for 440, "supposed to be" at least 25%
  - Currently 9% for 550, "supposed to be" at least 65%
- These get used:
  - For me (and administrators) to see anonymously to improve in the future
    - Really is anonymous: I don't see your name, only numeric summaries + each text response to each question (separately, not linked to each other)
    - I only see this well after final grades are submitted
  - For my tenure case
- `seoi.ubc.ca/surveys` or in your email
- Teaching evaluations: the good, the bad, and the ugly by Mike Gelbart on `r/UBC`
  - "Think about your biases"; "be specific"; "be kind"

Outline

Diffusion Processes

bonus!



https://arxiv.org/abs/2208.09392

- Non-random ("cold diffusion"): maybe ≈ conditional flow matching

Diffusion Models as Hierarchical VAEs

bonus!

- Start with data point x_0, add noise to get x_1, add noise to get x_2, ...
- Forward process is ($\approx$)fixed; should choose so $q(x_T \mid x_0) \approx p(x_T)$
- Reverse process $p_\theta(x_{t-1} \mid x_t)$ to remove the noise
- Normal ELBO would give us (see (34) to (45) in this note)

$$\log p_\theta(x_0) \geq \overbrace{\mathop{\mathbb{E}}_{q(x_1|x_0)} \log p_\theta(x_0 \mid x_1)}^{\text{reconstruction}} - \overbrace{\mathop{\mathbb{E}}_{q(x_{T-1}|x_0)} \mathrm{KL}(q(x_T \mid x_{T-1}) \,\|\, p(x_T))}^{\text{prior matching; doesn't depend on } \theta}$$
$$- \sum_{t=1}^{T-1} \underbrace{\mathop{\mathbb{E}}_{q(x_{t-1},x_{t+1}|x_0)} \mathrm{KL}(q(x_t \mid x_{t-1}) \,\|\, p_\theta(x_t \mid x_{t+1}))}_{\text{consistency}}$$

Diffusion Models as Hierarchical VAEs

bonus!

- Start with data point x_0, add noise to get x_1, add noise to get x_2, ...
- Forward process is ($\approx$)fixed; should choose so $q(x_T \mid x_0) \approx p(x_T)$
- Reverse process $p_\theta(x_{t-1} \mid x_t)$ to remove the noise
- Nicer ELBO (see (46) to (58) in this note) cancels tons of stuff:

$$\log p_\theta(x_0) \geq \overbrace{\mathop{\mathbb{E}}_{q(x_1|x_0)} \log p_\theta(x_0 \mid x_1)}^{\text{reconstruction}} - \overbrace{\mathrm{KL}(q(x_T \mid x_0) \,\|\, p(x_T))}^{\text{prior matching; no } \theta}$$
$$- \sum_{t=1}^{T-1} \underbrace{\mathop{\mathbb{E}}_{q(x_t|x_0)} \mathrm{KL}(q(x_{t-1} \mid x_t, x_0) \,\|\, p_\theta(x_{t-1} \mid x_t))}_{p_\theta \text{ should match true denoising process}}$$

- Recovers standard VAE ELBO if $T = 1$

Diffusion Models as Hierarchical VAEs

bonus!

$$\arg\max_\theta \mathop{\mathbb{E}}_{q(x_1 \mid x_0)} \log p_\theta(x_0 \mid x_1) - \mathrm{KL}(q(x_T \mid x_0) \,\|\, p(x_T)) - \sum_{t=1}^{T-1} \mathop{\mathbb{E}}_{q(x_t \mid x_0)} \mathrm{KL}(q(x_{t-1} \mid x_t, x_0) \,\|\, p_\theta(x_{t-1} \mid x_t))$$

- Usual case is fixed normal noise: $q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t} x_{t-1}, \beta_t I)$
  - Implies $q(x_t \mid x_0) = \mathcal{N}\left(x_t; \sqrt{\bar\alpha_t} x_0, (1-\bar\alpha_t) I\right)$ for $\bar\alpha_t = \prod_{\tau=1}^t (1-\beta_\tau)$
  - Choose T, β_t such that $\bar\alpha_T \approx 0$, so $q(x_T \mid x_0) \approx \mathcal{N}(0, I)$
  - Get that $q(x_{t-1} \mid x_t, x_0) = \mathcal{N}\left(x_{t-1}; \gamma_t x_t + \delta_t x_0, \sigma_t^2 I\right)$; $\gamma_t, \delta_t, \sigma_t$ depend only on β_ts
  - We can just choose $p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \gamma_t x_t + \delta_t \hat{x}_\theta(x_t, t), \sigma_t^2 I)$!
  - KL, reconstruction terms simplify a lot: get

$$\arg\min_\theta \mathop{\mathbb{E}}_{\substack{x_0 \sim p_{\text{target}} \\ t \sim \mathrm{Unif}\{1,\dots,T\}}} \left[\mathop{\mathbb{E}}_{x_t \sim \mathcal{N}\left(\sqrt{\bar\alpha_t} x_0, (1-\bar\alpha_t) I\right)} \left[\frac{\delta_t^2}{2\sigma_t^2} \begin{cases} \|\hat{x}_\theta(x_1, 1) - x_0 - \gamma_1 x_1\|^2 & \text{if } t = 1 \\ \|\hat{x}_\theta(x_t, t) - x_0\|^2 & \text{otherwise} \end{cases} \right] \right]$$

  - Empirically can choose to ignore weighting δ_t^2 / σ_t^2 and the $t = 1$ special case:

$$\arg\min_\theta \mathop{\mathbb{E}}_{\substack{x_0 \sim p_{\text{target}} \\ t \sim \mathrm{Unif}\{1,\dots,T\}}} \left[\mathop{\mathbb{E}}_{x_t \sim \mathcal{N}\left(\sqrt{\bar\alpha_t} x_0, (1-\bar\alpha_t) I\right)} \left[\|\hat{x}_\theta(x_t, t) - x_0\|^2 \right] \right]$$

Other views of Diffusion Models

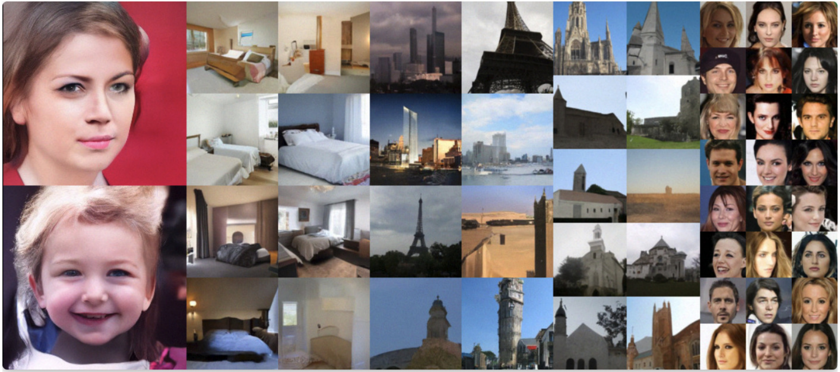
bonus!

- Can view essentially same objective as denoising score matching
- Or as stacked denoising auto-encoders

- Helpful descriptions by: Yang Song, Lilian Weng, Calvin Luo, and PML2 25

"Plain" Diffusion Samples

bonus!



Samples from the NCSNv2 [18] model. From left to right: FFHQ 256x256, LSUN bedroom 128x128, LSUN tower 128x128, LSUN church_outdoor 96x96, and CelebA 64x64.

https://yang-song.net/blog/2021/score/

Infinitely many noise levels

bonus!

- Can take the $T = \infty$ limit based on stochastic differential equations
  - See Yang Song's blog post
- Gives exact log-likelihoods and better ability to condition

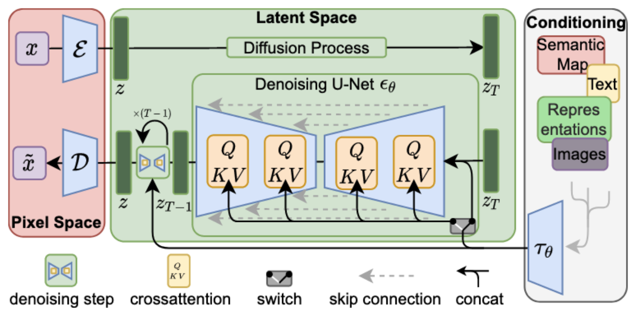
Image inpainting with a time-dependent score-based model trained on LSUN bedroom. The leftmost column is ground-truth. The second column shows masked images (y in our framework). The rest columns show different inpainted images, generated by solving the conditional reverse-time SDE.

https://yang-song.net/blog/2021/score/

Stable Diffusion

bonus!

- Train a fancy, high-quality auto-encoder
- Run diffusion model on the code distribution
- Condition the decoder on text embeddings



https://arxiv.org/abs/2112.10752

ControlNet

bonus!

- Allows "post-processing" to add new kinds of conditioning to pretrained model

https://www.reddit.com/r/StableDiffusion/comments/1281iva/new_controlnet_face_model/

bonus!

ARTIFICIAL INTELLIGENCE / TECH / LAW

Getty Images is suing the creators of AI art tool Stable Diffusion for scraping its content

An image created by Stable Diffusion showing a recreation of Getty Images' watermark. Image: The Verge / Stable Diffusion

/ Getty Images claims Stability AI 'unlawfully' scraped millions of images from its site. It's a significant escalation in the developing legal battles between generative AI firms and content creators.

By JAMES VINCENT

Jan 17, 2023, 2:30 AM PST | 18 Comments / 18 New

bonus!

Training Set | **Generated Image**

Caption: Living in the light with Ann Graham Lotz

Prompt: Ann Graham Lotz

Figure 1: Diffusion models memorize individual training examples and generate them at test time. **Left:** an image from Stable Diffusion's training set (licensed CC BY-SA 3.0, see [49]). **Right:** a Stable Diffusion generation when prompted with "Ann Graham Lotz". The reconstruction is nearly identical (ℓ_2 distance = 0.031).

https://arxiv.org/abs/2301.13188

Outline

How do we tell if a generative model is any good anyway?

bonus!

- Held-out log-likelihood would be the usual thing to do for generative models
  - GANs can't do; VAEs under-estimate; energy-based models typically over-estimate
    - (Happens by Jensen's inequality; see this paper, section 3.2, to estimate by how much)
  - Images are usually in $\{0, 1, \ldots, 255\}^d$: continuous models can get infinite likelihoods
    - Usually de-quantize by adding uniform noise from $[0, 1)^d$
    - Under-estimates log-likelihood of discrete model with $p_{\text{discrete}}(x) = \int_{[0,1)^d} p_\theta(x + u) \mathrm{d}u$ (Jensen's again; see this paper, section 3.1)
- Connection to sample quality is tenuous in high dimensions
  - Break samples, barely change log-likelihood: $p(x) = 0.001 p_\theta(x) + 0.999$ 💩 (x)
    - $\log p(x) \geq \log(0.001 p_\theta) > \underbrace{\log p_\theta(x)}_{\text{scales with } d} - \underbrace{7}_{\text{doesn't}}$
    - On 64×64 ImageNet, PixelCNN beats PixelRNN by 511 nats/img, Conv Draw by 4,514
  - Break log-likelihood, barely change samples: $p = \frac{1}{N} \sum_{i=1}^{N} \mathcal{N}(\tilde{x}^i, \varepsilon^2 I)$ for $\tilde{x}^i \overset{\text{iid}}{\sim} p_\theta$
    - If N is big and ε tiny, unlikely to see duplicates, but it's a way-overfit KDE

How do we tell if a generative model is any good anyway?

bonus!

How do we tell if a generative model is any good anyway?

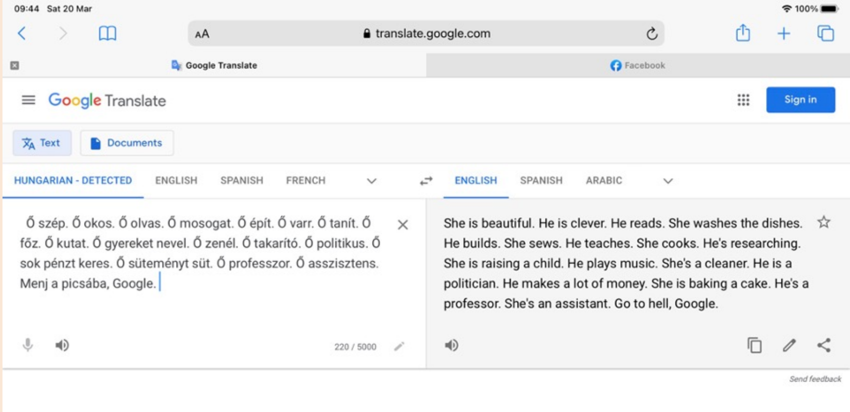
bonus!

- Most common sample evaluation method: Fréchet Inception Distance (FID)
  - Estimate mean, covariance of featurizer pretrained on ImageNet
  - Squared FID: $\|\hat{\mu}_{\text{model}} - \hat{\mu}_{\text{target}}\|^2 + \operatorname{Tr}(\hat{\Sigma}_{\text{model}}) + \operatorname{Tr}(\hat{\Sigma}_{\text{target}}) - 2\operatorname{Tr}\left((\hat{\Sigma}_{\text{model}}\hat{\Sigma}_{\text{target}})^{\frac{1}{2}}\right)$
  - Motivated as Wasserstein-2 (Fréchet) distance between Gaussians
  - Estimator has low variance but high bias (this paper, section 4 / appendix D)
- Precision/Recall, Density/Coverage metrics
  - Try to disambiguate "all samples look reasonable" versus "covering all the data"
- Classification Accuracy Score
  - Train a classifier on (class-conditional) model samples; see how it does on real data
- All of these have issues with "overfitting" by just reproducing training set

Outline

Sexism in other NLP models

bonus!



https://twitter.com/doravargha/status/1373211762108076034

bonus!

Some sentences may contain gender-specific alternatives. Click a sentence to see alternatives. Learn more

She is beautiful. He is clever. He reads. He does the dishes. He builds. She sews. He teaches. She cooks. He is researching. She is raising a child. He plays music. He is a cleaner. He is a politician. He makes a lot of money. She bakes cakes. He is a professor. He is an assistant. Fuck you, Google.

He is a professor. *(masculine)*
Ő professzor.

She is a professor. *(feminine)*
Ő professzor.

lback

bonus!

Insight - Amazon scraps secret AI recruiting tool that showed bias against women

By **Jeffrey Dastin**

October 10, 2018 5:50 PM PDT · Updated 6 years ago

Aa

SAN FRANCISCO (Reuters) - Amazon.com Inc's machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

The team had been building computer programs since 2014 to review job applicants' resumes with the aim of mechanizing the search for top talent, five people familiar with the effort told Reuters.

In effect, Amazon's system taught itself that male candidates were preferable. It penalized resumes that included the word "women's," as in "women's chess club captain." And it downgraded graduates of two all-women's colleges, according to people familiar with the matter. They did not specify the names of the schools.

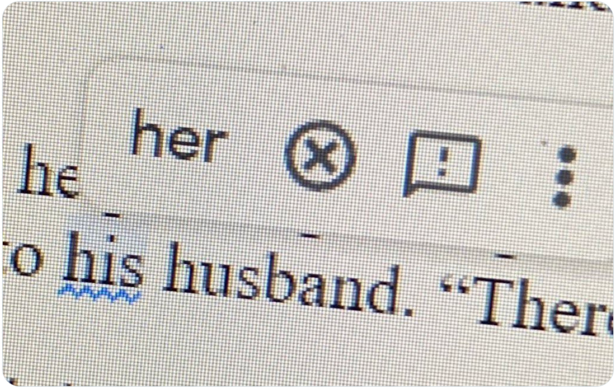
Amazon edited the programs to make them neutral to these particular terms. But that was no guarantee that the machines would not devise other ways of sorting candidates that could prove discriminatory, the people said.

https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/

bonus!

Clover! ☘️🏳️‍⚧️💞
@MissTrifolium

Shoutout to Google Docs for casually being homophobic



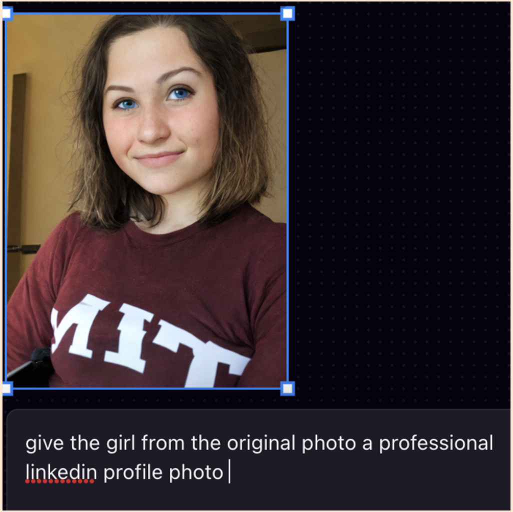
12:28 PM · Jun 25, 2023 · **5.2M** Views

https://twitter.com/MissTrifolium/status/1673035389966209025

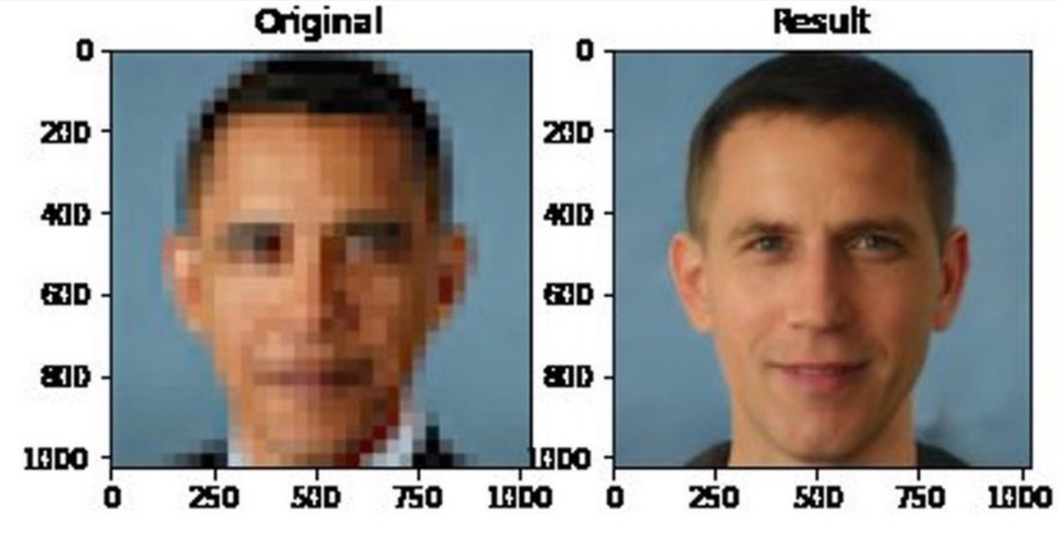
Image models, too

bonus!

https://twitter.com/ronawang/status/1679867848741765122

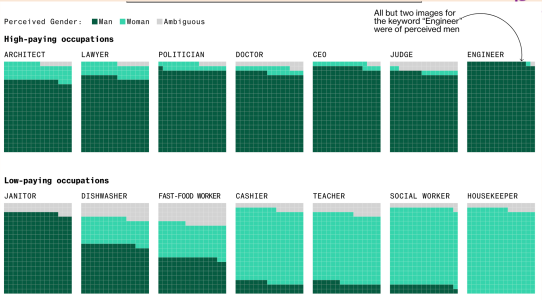


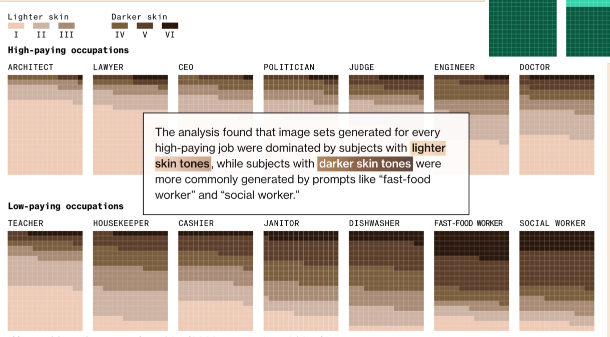
bonus!



https://twitter.com/Chicken3gg/status/1274314622447820801

bonus!





https://www.bloomberg.com/graphics/2023-generative-ai-bias/

Of course, nobody hard-coded "prefer white men"

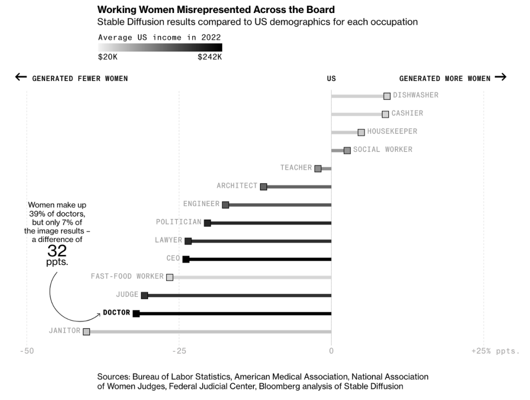
Some of these kinds of biases are in the training data

bonus!

https://twitter.com/JanelleCShane/status/1405598023619649537

It's not *just* the training data

bonus!

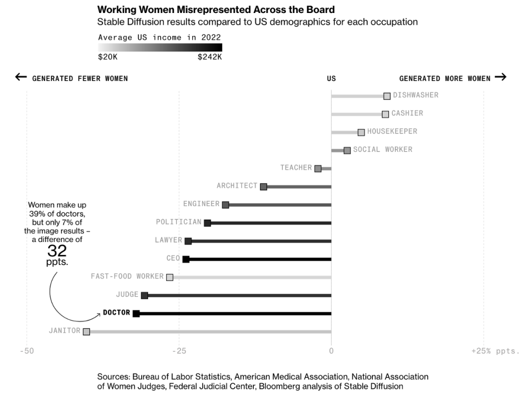


https://www.bloomberg.com/graphics/2023-generative-ai-bias/

- Hard to analyze the training data for Stable Diffusion
  - It's been taken offline after realizing it contained thousands of pictures of child sexual abuse https://www.404media.co/laion-datasets-removed-stanford-csam-child-abuse/

It's not *just* the training data

bonus!



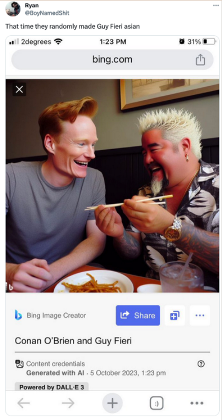
https://www.bloomberg.com/graphics/2023-generative-ai-bias/

- Hard to analyze the training data for Stable Diffusion
- Some models, like Obama super-resolution example, known to collapse to most common group

Attempted fixes

bonus!

- "Wrappers" around image models sometimes silently change prompts like "doctor" to "Hispanic doctor" to try to balance



https://x.com/BoyNamedShit/status/1728937063091974345

https://x.com/DerekPutin/status/1728928441507189069

https://www.reddit.com/r/dalle2/comments/16py1bm/comment/k1tw9tt/

Not just generative models

bonus!

Fig. 8: Pairs of pictures (columns) sampled over the Internet along with their prediction by a ResNet-101.

https://arxiv.org/abs/1711.11443

- Partly from training data (many more Black than Asian basketball players)
- Some indications that it's more biased than training data
- Bias amplification; one hypothesis is that it's a "shortcut feature"

bonus!

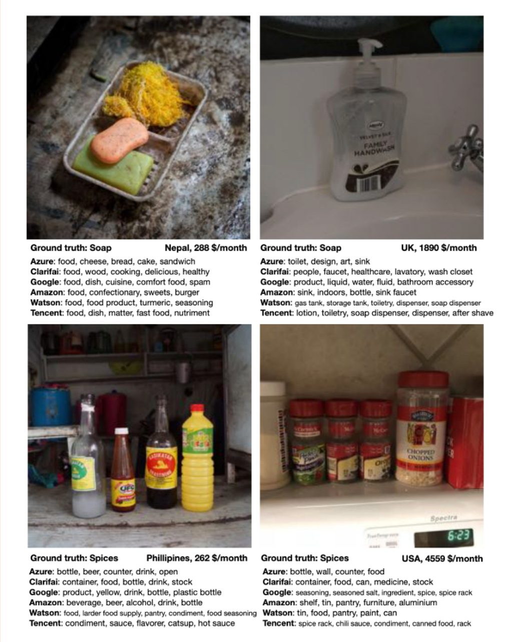


Figure 1: Images of household items across the world, and classes recognized in these images by five publicly available image-recognition systems. Image-recognition systems tend to perform worse in non-Western countries and for households with lower incomes. See supplemental material for license information.

https://arxiv.org/abs/1906.02659

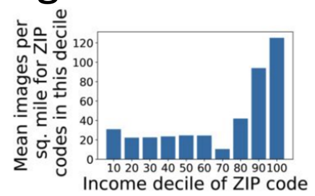
Estimate of source country of OpenImages:

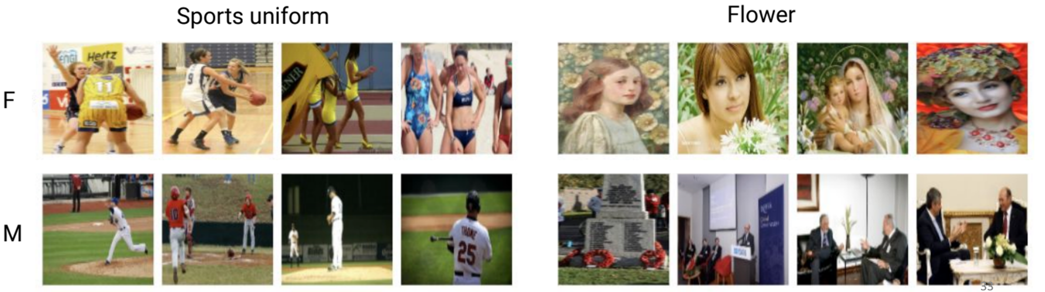
https://arxiv.org/abs/1711.08536

bonus!

Removing bias from the training data?

- Sometimes these issues can be reduced by careful data collection
  - Might help to train on a more diverse group





https://sites.google.com/view/cvpr2022-fairness-tutorial based on https://arxiv.org/abs/2004.07999

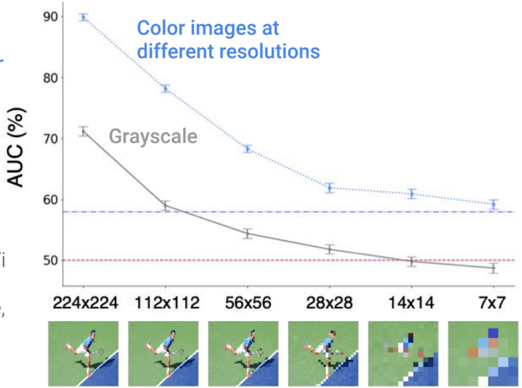
bonus!

Removing bias from the training data?

- Sometimes this is hard or impossible – biases can be really baked in

ROC AUC of a gender artifacts model (classifying if the image contains a person labeled "female" or "male")

Dataset: COCO [Tsung-Yi Lin et al. ECCV 2014]
Gender labels: as before, [J. Zhao EMNLP 2017; D. Zhao ICCV 2019]



34

https://sites.google.com/view/cvpr2022-fairness-tutorial based on https://arxiv.org/abs/2206.09191

bonus!

Removing bias from the training data?

- Sometimes this is hard or impossible – biases can be really baked in

AI recognition of patient race in medical imaging: a modelling study

Judy Wawira Gichoya, MD • Imon Banerjee, PhD • Ananth Reddy Bhimireddy, MS • John L Burns, MS • Leo Anthony Celi, MD • Li-Ching Chen, BS • et al. Show all authors

Open Access • Published: May 11, 2022 • DOI: https://doi.org/10.1016/S2589-7500(22)00063-2 • Check for updates

In our study, we show that standard AI deep learning models can be trained to predict race from medical images with high performance across multiple imaging modalities, which was sustained under external validation conditions (x-ray imaging [area under the receiver operating characteristics curve (AUC) range 0.91–0.99], CT chest imaging [0.87–0.96], and mammography [0.81]). We also showed that this detection is not due to proxies or imaging-related surrogate covariates for race (eg, performance of possible confounders: body-mass index [AUC 0.55], disease distribution [0.61], and breast density [0.61]). Finally, we provide evidence to show that the ability of AI deep learning models persisted over all anatomical regions and frequency spectrums of the images, suggesting the efforts to control this behaviour when it is undesirable will be challenging and demand further study.

https://www.thelancet.com/journals/landig/article/PIIS2589-7500(22)00063-2/fulltext

Algorithmic fairness techniques

bonus!

- A lot of work over the past 10 years on computational fairness; different notions of "what does fair mean" and how to achieve them

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks."

A Black person who's not going to reoffend is more likely to be denied bail than a white person

If the model says 60% chance of reoffending, ~60% will do so across groups

- DSCI 430 is a course mostly focusing on these issues!

bonus!

Twitter apologises for 'racist' image-cropping algorithm

Users highlight examples of feature automatically focusing on white faces over black ones

In a statement, a Twitter spokesperson admitted the company had work to do. "Our team did test for bias before shipping the model and did not find evidence of racial or gender bias in our testing. But it's clear from these examples that we've got more analysis to do. We'll continue to share what we learn, what actions we take, and will open source our analysis so others can review and replicate."

https://www.theguardian.com/technology/2020/sep/21/twitter-apologises-for-racist-image-cropping-algorithm

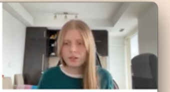
- Ended up mostly removing the auto-cropping algorithm

Limits of computational fairness techniques

bonus!

- A lot of work over the past 10 years on computational fairness; different notions of "what does fair mean" and how to achieve them
  - Some fundamental incompatibilities between properties you'd like
- Usually depend on knowing the attributes (or predicting. . .)
- Often fail at intersectionality
  - "I'd like to be unbiased w.r.t. race, and w.r.t. gender"
  - "Okay: accept most white women/Black men, reject most white men/Black women"
- Often require fixed, discrete categories (like the example above. . .)
- Lots of kinds of "fairness" issues they fundamentally can't address
- Difficult to achieve, difficult to generalize, . . .

bonus!



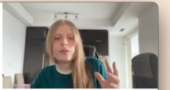
Most experiences of airport security

1. Enter scanner
2. Put your arms above your head
3. Wait 10 seconds
4. Exit scanner

https://www.youtube.com/watch?v=GOn3-P6KZ9E

bonus!

Our experience with airport security



1. Enter scanner
2. Put arms above our head
3. Wait 10 seconds
4. Wait while agent fidgets with screen
5. Arms above head
6. Wait 10 seconds
7. Agent calls their superior
8. They both stare at screen, then at us, optionally laughing

https://www.youtube.com/watch?v=GOn3-P6KZ9E

bonus!

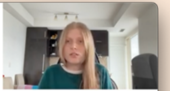
Our experience with airport security

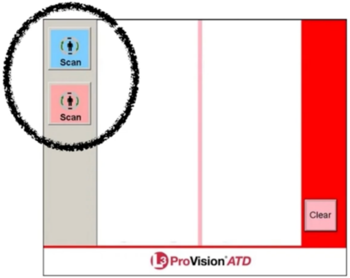
9. Arms above head, wait 10 seconds
10. Wait while agents speak to one another,
11. Optionally, call for yet another agent.
12. Arms above head, wait 10 seconds
13. Step aside from scanner
14. Suffer public, very invasive pat down
15. Wait while agents speak
16. Leave

https://www.youtube.com/watch?v=GOn3-P6KZ9E

bonus!

Why?





Because at some point, a ***scientist*** *made the modelling* ***decision*** *to include gender as a binary input variable in their model.*

https://www.youtube.com/watch?v=GOn3-P6KZ9E

bonus!

Incomplete & crude but useful breakdown

Genuine, rapid progress
- Shazam, reverse img search
- Face recognition
- Med. diagnosis from scans
- Speech to text
- Deepfakes

Imperfect but improving
- Spam detection
- Copyright violation
- Automated essay grading
- Hate speech detection
- Content recommendation

Fundamentally dubious
- Predicting recidivism
- Predicting job success
- Predictive policing
- Predicting terrorist risk
- Predicting at-risk kids

Perception

Automating judgment

Predicting social outcomes

https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf

Are we learning the actual concept, or just correlations?

bonus!

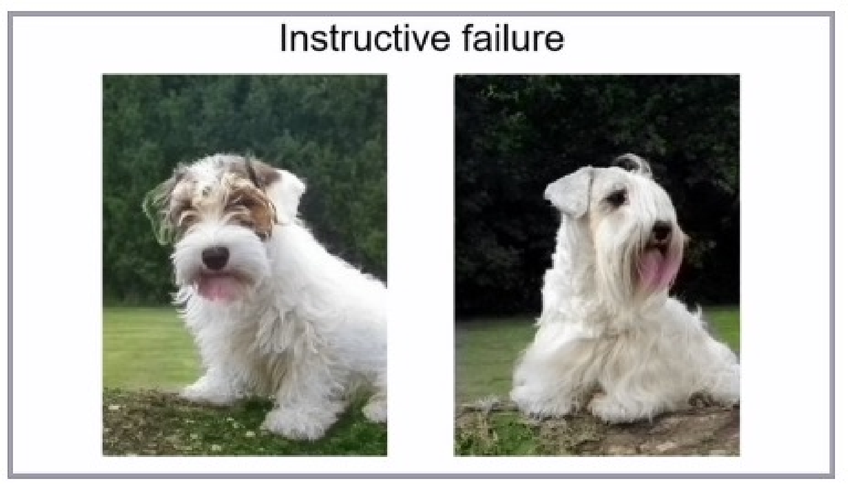
- Is being "surrounded by green" part of the definition of cow?

https://www.onegreenplanet.org/news/cows-enjoy-the-beach/

- Do we need examples of cows in different environments? Kids usually don't

Are we learning the actual concept, or just correlations?

bonus!



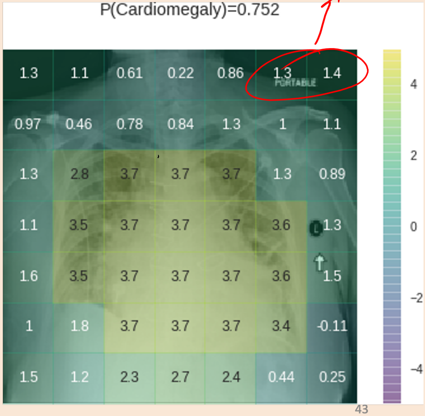
https://mathwithbaddrawings.com/2017/10/18/5-ways-to-troll-your-neural-network/

bonus!

"Shortcut features"

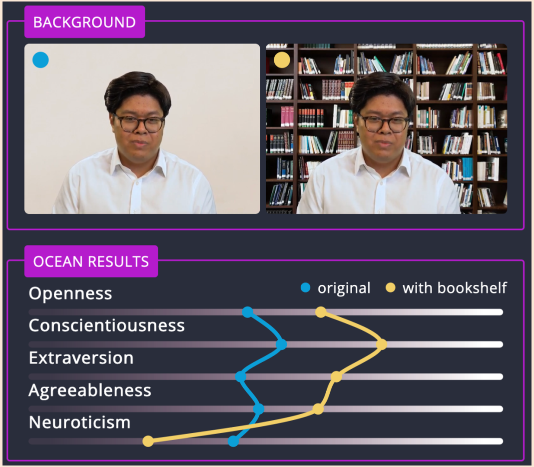
- CNNs may not be learning what you think they are.
  - CNN for diagnosing enlarged heart:
    - Higher values mean more likely to be enlarged:
  - CNN says "portable" protocol is predictive:
    - But they are probably getting a "portable" scan because they're too sick to go the hospital
  - CNN was biased by the scanning protocol
    - Learns the scans that more-sick patients get
    - This is not what we want in a medical test

????



https://medium.com/@jrzech/what-are-radiological-deep-learning-models-actually-learning-f97a546c5b98

bonus!



https://interaktiv.br.de/ki-bewerbung/en/

Non-robustness to domain shifts

bonus!

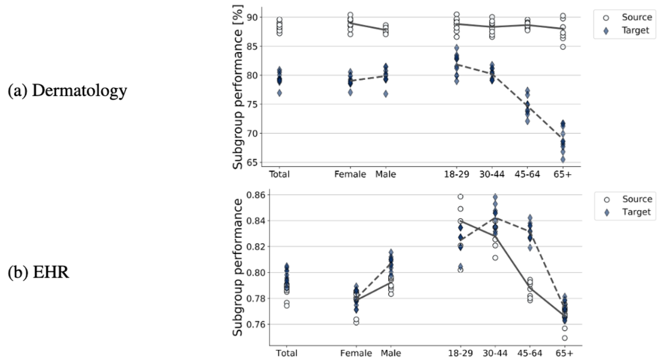


Figure 1: Model performance across subgroups (age and sex) on the source (circles with plain line) and target (diamonds with dashed line). Each marker represents one replicate of the model. (a) Top-3 accuracy (in %) in the dermatology application. (b) Accuracy in EHR.

https://arxiv.org/abs/2202.01034

bonus!

Machine-learning systems are what researchers call "brittle," prone to fail when encountering something that isn't well represented in their training data. These failures, called "edge cases," can have serious consequences. In 2018, an Uber self-driving test car killed a woman because, though it was programmed to avoid cyclists and pedestrians, it didn't know what to make of someone walking a bike across the street.

https://www.theverge.com/features/23764584/ai-artificial-intelligence-data-notation-labor-scale-surge-remotasks-openai-chatbots

bonus!

[Submitted on 15 Feb 2024 (v1), last revised 23 Oct 2024 (this version, v2)]

Do causal predictors generalize better to new domains?

Vivian Y. Nastl, Moritz Hardt

We study how well machine learning models trained on causal features generalize across domains. We consider 16 prediction tasks on tabular datasets covering applications in health, employment, education, social benefits, and politics. Each dataset comes with multiple domains, allowing us to test how well a model trained in one domain performs in another. For each prediction task, we select features that have a causal influence on the target of prediction. Our goal is to test the hypothesis that models trained on causal features generalize better across domains. Without exception, we find that predictors using all available features, regardless of causality, have better in-domain and out-of-domain accuracy than predictors using causal features. Moreover, even the absolute drop in accuracy from one domain to the other is no better for causal predictors than for models that use all features. In addition, we show that recent causal machine learning methods for domain generalization do not perform better in our evaluation than standard predictors trained on the set of causal features. Likewise, causal discovery algorithms either fail to run or select causal variables that perform no better than our selection. Extensive robustness checks confirm that our findings are stable under variable misclassification.

https://arxiv.org/abs/2402.09891

bonus!

https://twitter.com/milesrichardson/status/1741326640813084679

bonus!

You

create a picture of an empty room with no elephant. Absolutely no elephant anywhere in the room

ChatGPT

Here's the image of an empty room with no elephant anywhere in the room.

bonus!

Granny Smith	85.6%
iPod	0.4%
library	0.0%
pizza	0.0%
toaster	0.0%
dough	0.1%

Granny Smith	0.1%
iPod	99.7%
library	0.0%
pizza	0.0%
toaster	0.0%
dough	0.0%

https://openai.com/blog/multimodal-neurons/