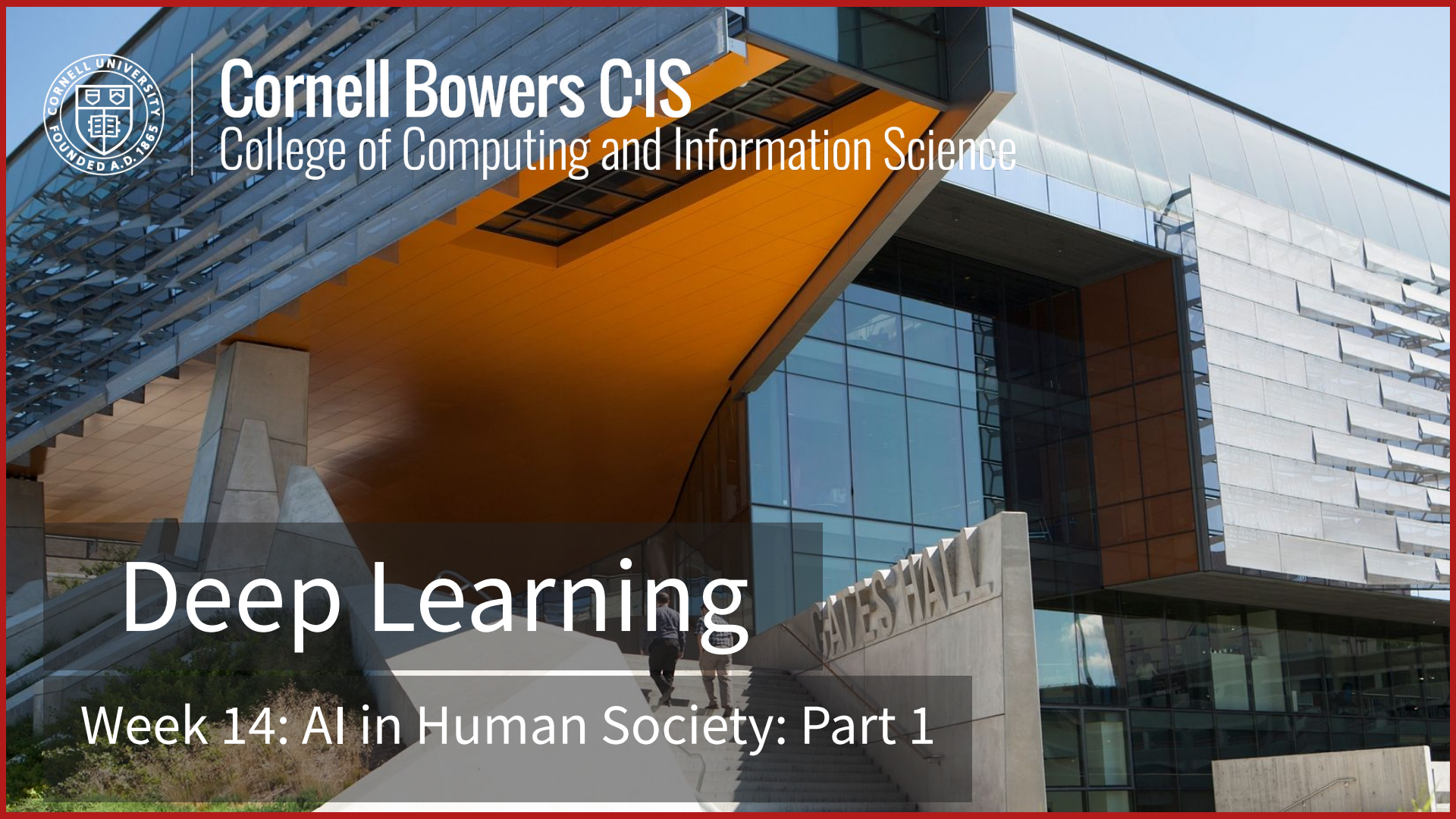


Deep Learning

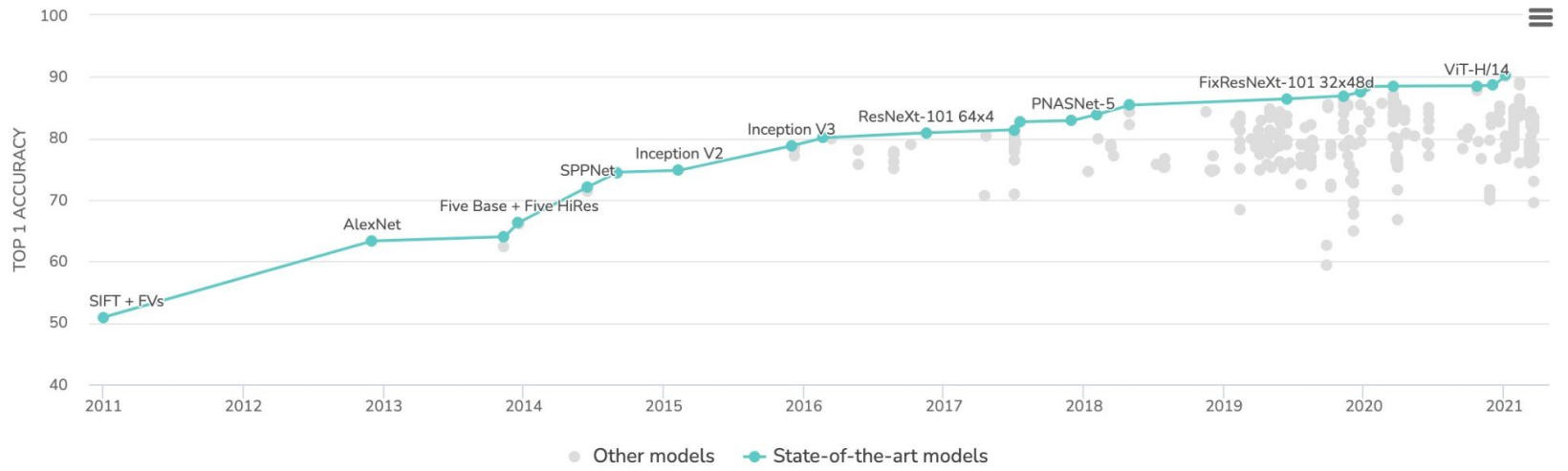
Week 14: AI in Human Society: Part 1

Robustness

- How well do models generalize to real data?
  - IID assumption often doesn't hold in the real world

Image Classification on ImageNet

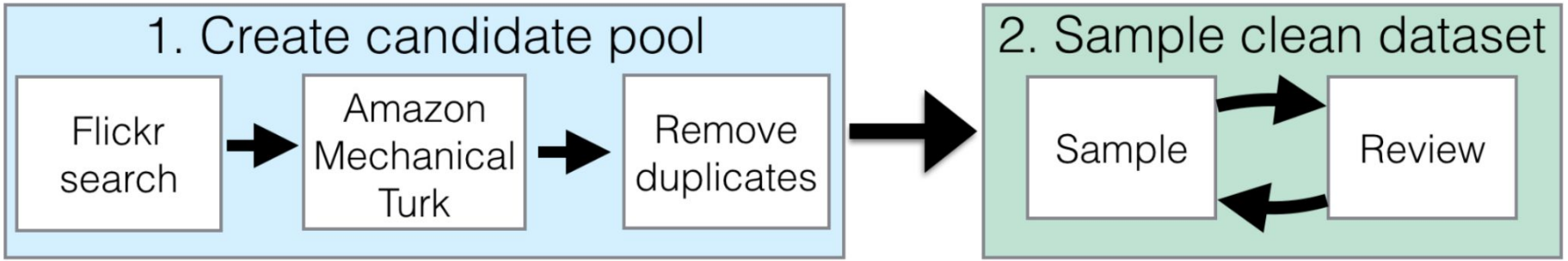
Leaderboard Dataset



https://paperswithcode.com/sota/image-classification-on-imagenet

Do ImageNet Classifiers Generalize to ImageNet?

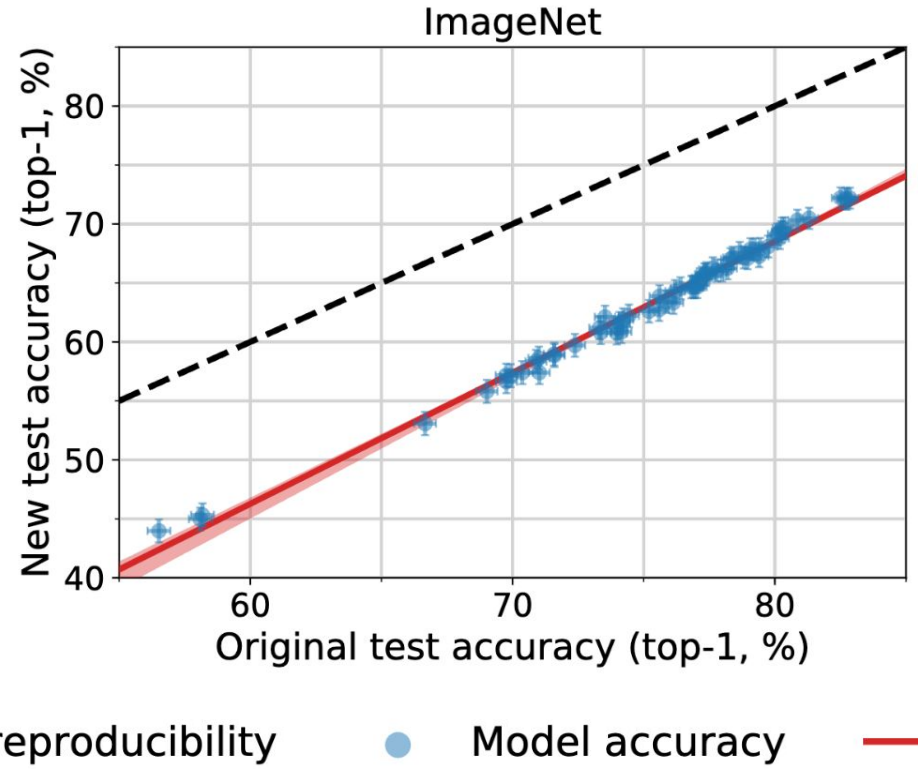
- Recreate the ImageNet test set following the original data curation protocol
  - Query images uploaded to Flickr in the same time frame



Recht, Benjamin, et al. "Do imagenet classifiers generalize to imagenet?." International conference on machine learning. PMLR, 2019.

Do ImageNet Classifiers Generalize to ImageNet?

- Significant performance drop across all models
- Hard to satisfy the IID assumption
- Model performance lies on a line
  - Ranking largely unchanged!

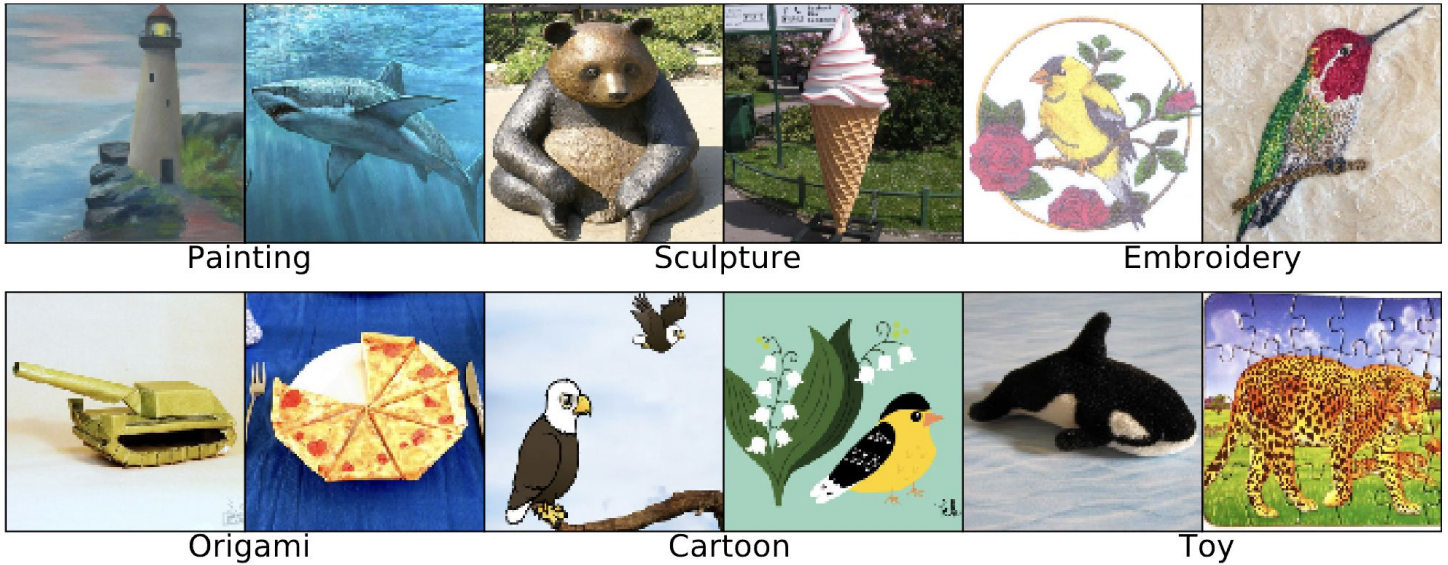


Recht, Benjamin, et al. "Do imagenet classifiers generalize to imagenet?." International conference on machine learning. PMLR, 2019.

Robustness Under Distribution Shift

ImageNet-Renditions

- Collection of images for 200 ImageNet classes in various different styles



Hendrycks, Dan, et al. "The many faces of robustness: A critical analysis of out-of-distribution generalization." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

Robustness Under Distribution Shift

ImageNet-Renditions

- Top1-Error increases dramatically
  - 6.8% → 58.7%

	ImageNet-200 (%)	ImageNet-R (%)	Gap
ResNet-50	7.9	63.9	56.0
ResNet-152 (*Larger Models*)	6.8	58.7	51.9

Robustness Under Distribution Shift

ImageNet-Sketch

- Collection of black-and-white sketches for ImageNet classes

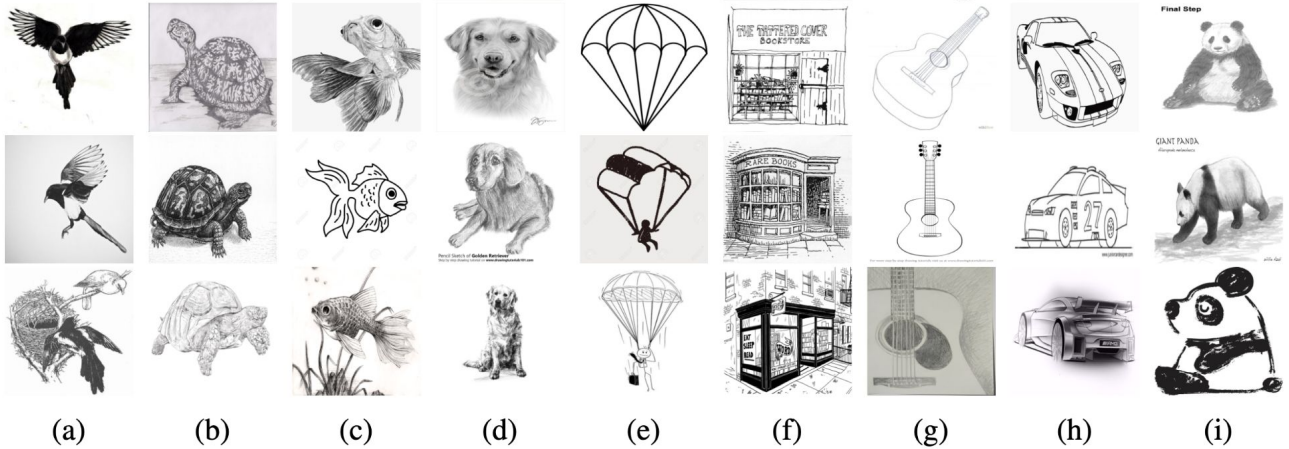


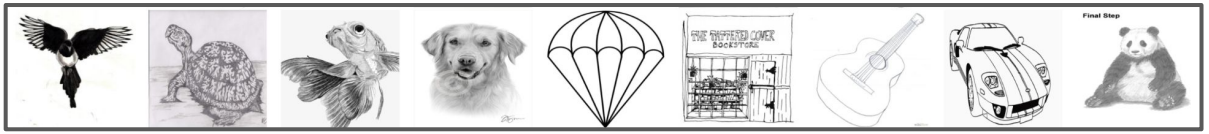
Figure 3: Sample Images from ImageNet-Sketch. Corresponding classes: (a) magpie (b) box turtle (c) goldfish (d) golden retriever (e) parachute (f) bookshop (g) acoustic guitar (h) racer (i) giant panda

Wang, Haohan, et al. "Learning robust global representations by penalizing local predictive power." Advances in Neural Information Processing Systems 32 (2019).

Cornell Bowers CIS

Robustness Under Distribution Shift

ImageNet-Sketch



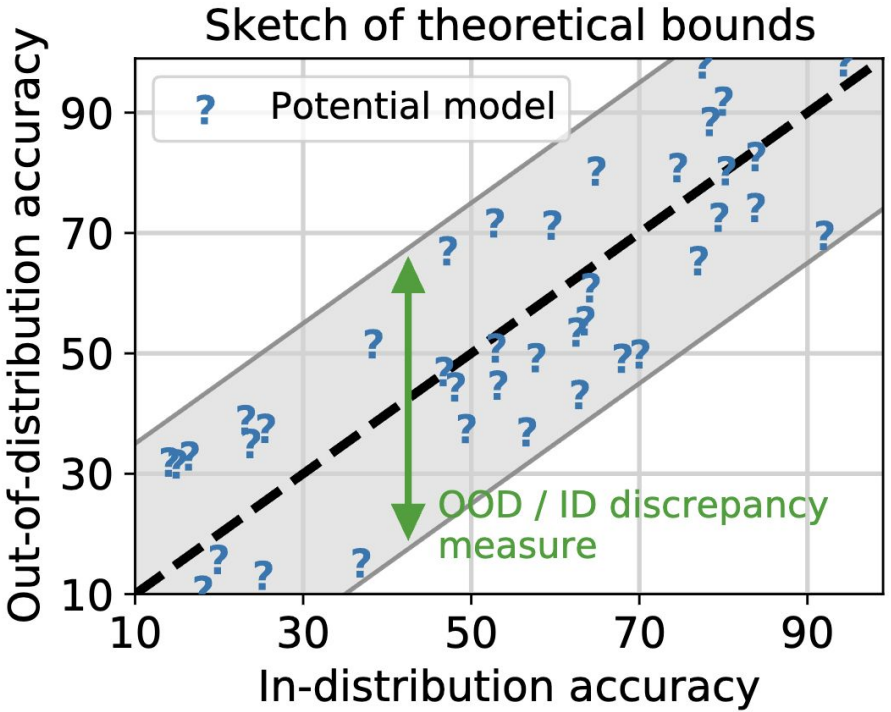
- Significant drop in accuracy
- Data augmentation improves performance
  - Never approach in-domain performance

Augmentation(s)	ImageNet		IN-Sketch	
	top-1	top-5	top-1	top-5
Baseline	76.6%	**93.3%**	22.4%	39.3%
+ Color distortion	**76.9%**	**93.3%**	28.1%	46.6%
+ Gaussian blur	76.8%	**93.3%**	29.0%	47.9%
+ Gaussian noise	75.9%	92.8%	29.8%	48.9%
+ Min. crop of 64%	73.5%	91.5%	**30.9%**	**51.4%**
+ Stronger aug.	72.0%	90.7%	30.4%	50.5%
+ Longer training	71.1%	90.0%	30.5%	50.4%

Wang, Haohan, et al. "Learning robust global representations by penalizing local predictive power." Advances in Neural Information Processing Systems 32 (2019).

Accuracy on the Line

How does OOD performance relate to ID performance?

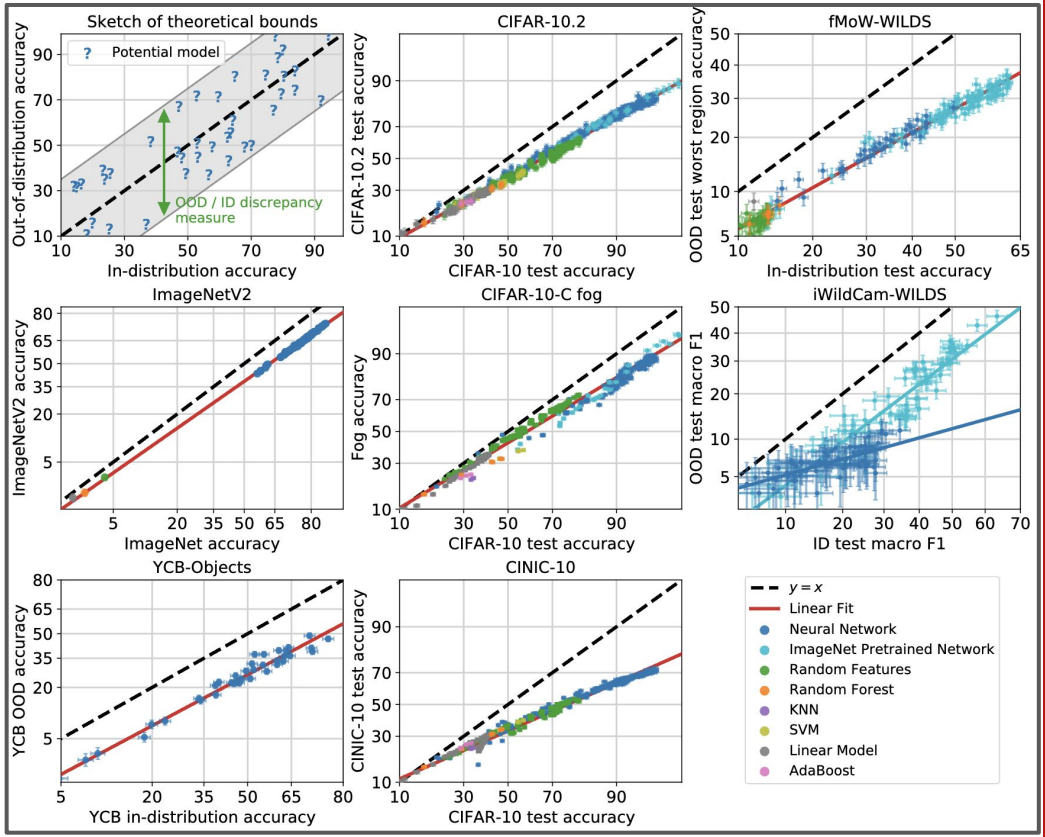


Cornell Bowers CIS

Accuracy on the Line

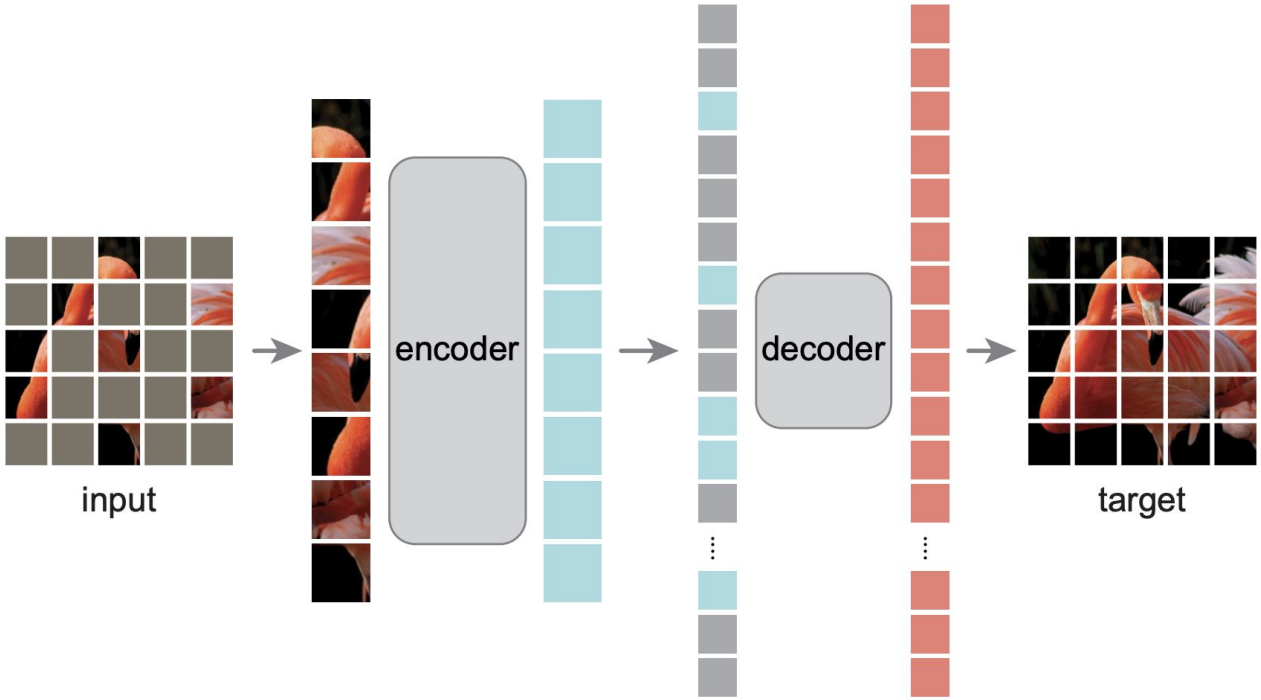
OOD accuracy is often linearly correlated with ID accuracy

- OOD performance scales predictably with ID performance
- Ranking of models is stable across settings

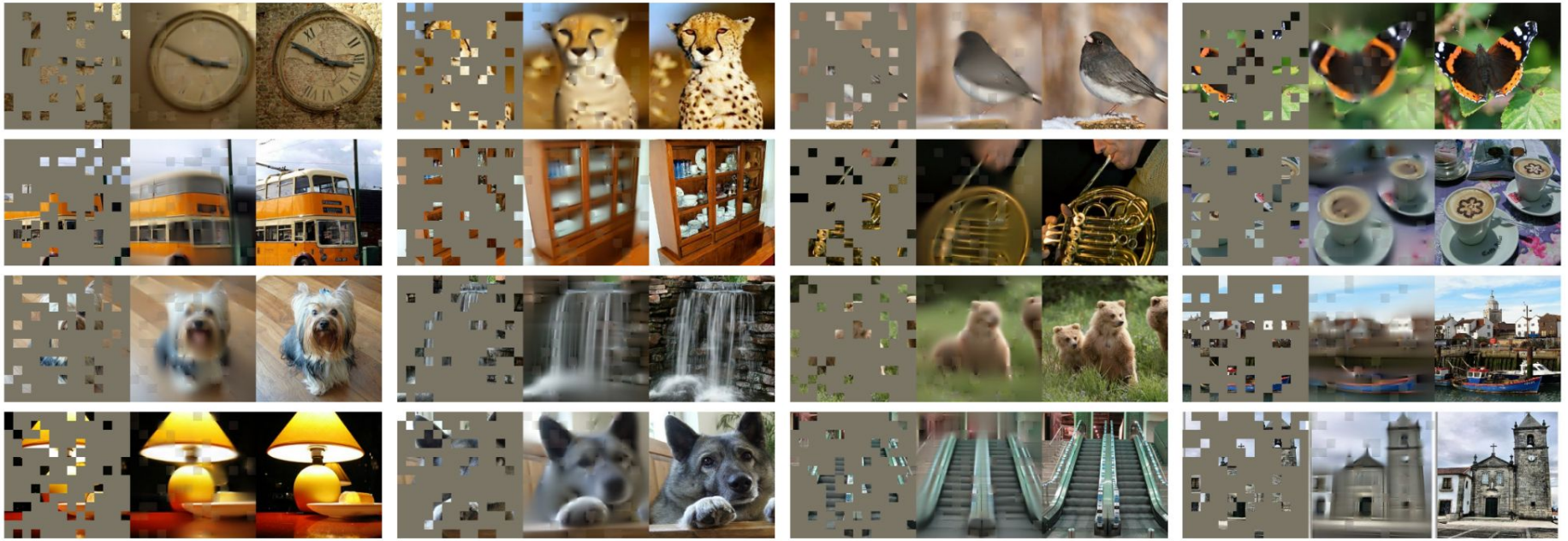


Miller, John P., et al. "Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization." International conference on machine learning. PMLR, 2021.

Self-Supervised Learning: Masked Autoencoders (MAE)



He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16000-16009).



Cornell Bowers CIS

MAE Robustness

- Compared to supervised ViTs
  - Outperforms on in-domain data
  - Significantly outperforms on out-of-distribution data

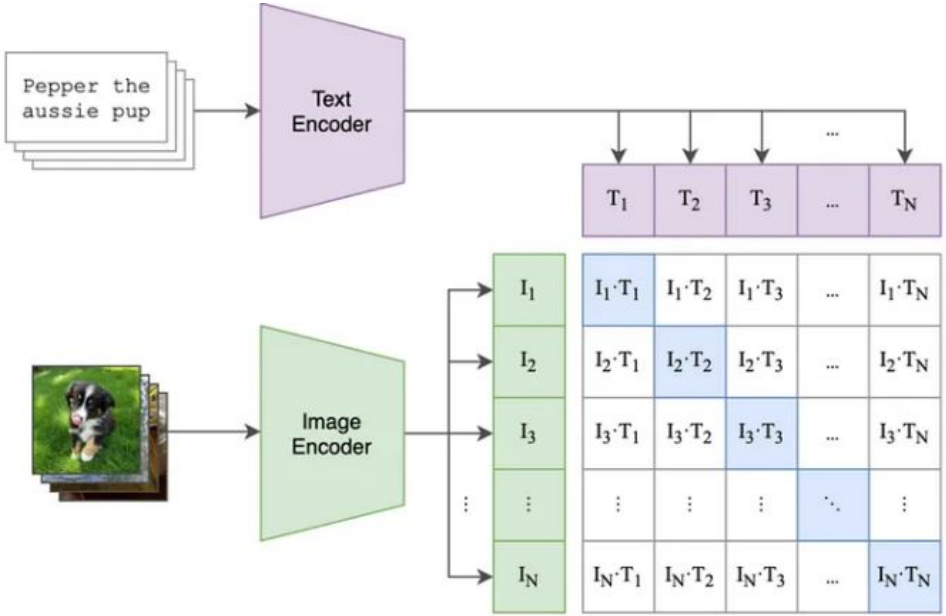
method	pre-train data	ViT-B	ViT-L	ViT-H	ViT-H_{448}
scratch, our impl.	-	82.3	82.6	83.1	-
DINO [5]	IN1K	82.8	-	-	-
MoCo v3 [9]	IN1K	83.2	84.1	-	-
BEiT [2]	IN1K+DALLE	83.2	85.2	-	-
MAE	IN1K	83.6	85.9	86.9	**87.8**

In-Domain Performance

dataset	ViT-B	ViT-L	ViT-H	ViT-H_{448}	prev best
IN-Corruption ↓ [27]	51.7	41.8	**33.8**	36.8	42.5 [32]
IN-Adversarial [28]	35.9	57.1	68.2	**76.7**	35.8 [41]
IN-Rendition [26]	48.3	59.9	64.4	**66.5**	48.7 [41]
IN-Sketch [60]	34.5	45.3	49.6	**50.9**	36.0 [41]
our supervised training baselines:					
IN-Corruption ↓	45.8	42.3	**41.3**		
IN-Adversarial	27.2	29.6	**33.1**		
IN-Rendition	49.4	**50.9**	50.3		
IN-Sketch	35.6	37.5	**38.0**		

Out-of-Domain Performance

CLIP (Contrastive Language–Image Pre-training)



Conde, M. V., & Turgutlu, K. (2021). CLIP-Art: Contrastive pre-training for fine-grained art classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3956-3960).

FOOD101

guacamole (90.1%) Ranked 1 out of 101 labels

- ✓ a photo of **guacamole**, a type of food.
- × a photo of **ceviche**, a type of food.
- × a photo of **edamame**, a type of food.
- × a photo of **tuna tartare**, a type of food.
- × a photo of **hummus**, a type of food.

SUN397

television studio (90.2%) Ranked 1 out of 397

- ✓ a photo of a **television studio**.
- × a photo of a **podium indoor**.
- × a photo of a **conference room**.
- × a photo of a **lecture room**.
- × a photo of a **control room**.

YOUTUBE-BB

airplane, person (89.0%) Ranked 1 out of 23

- ✓ a photo of a **airplane**.
- × a photo of a **bird**.
- × a photo of a **bear**.
- × a photo of a **giraffe**.
- × a photo of a **car**.

EUROSAT

annual crop land (12.9%) Ranked 4 out of 10

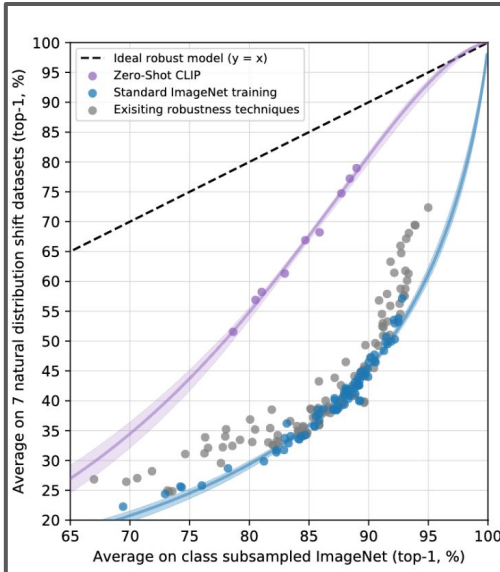
- × a centered satellite photo of **permanent crop land**.
- × a centered satellite photo of **pasture land**.
- × a centered satellite photo of **highway or road**.
- ✓ a centered satellite photo of **annual crop land**.
- × a centered satellite photo of **brushland or shrubland**.

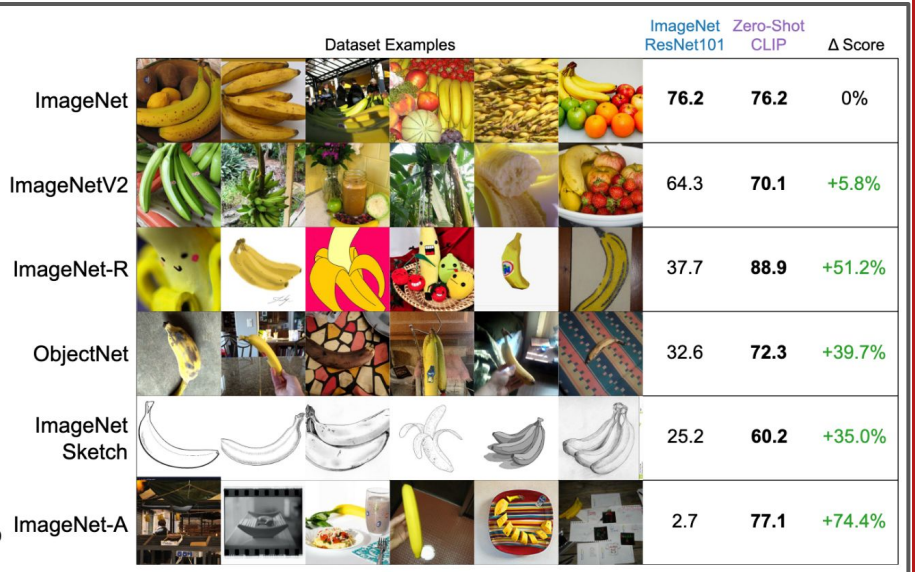
https://openai.com/research/clip

CLIP Robustness

Significantly improved robustness over supervised ImageNet training

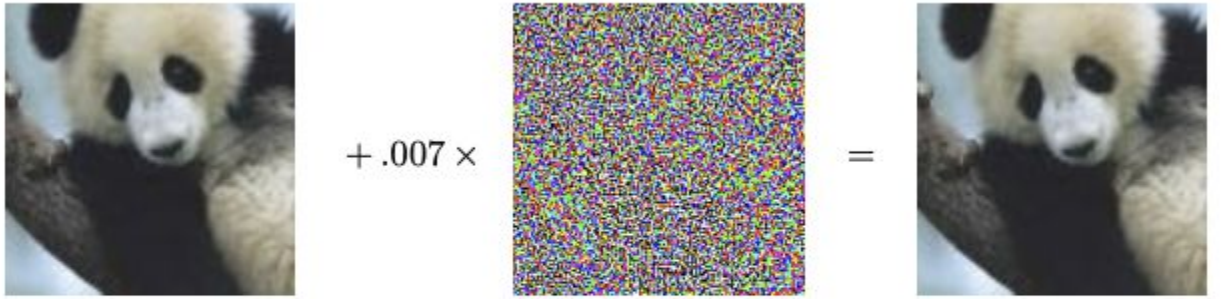
- No longer “out-of-distribution”
  - CLIP pre-training data likely has paintings, sketches, etc.





Adversarial Examples

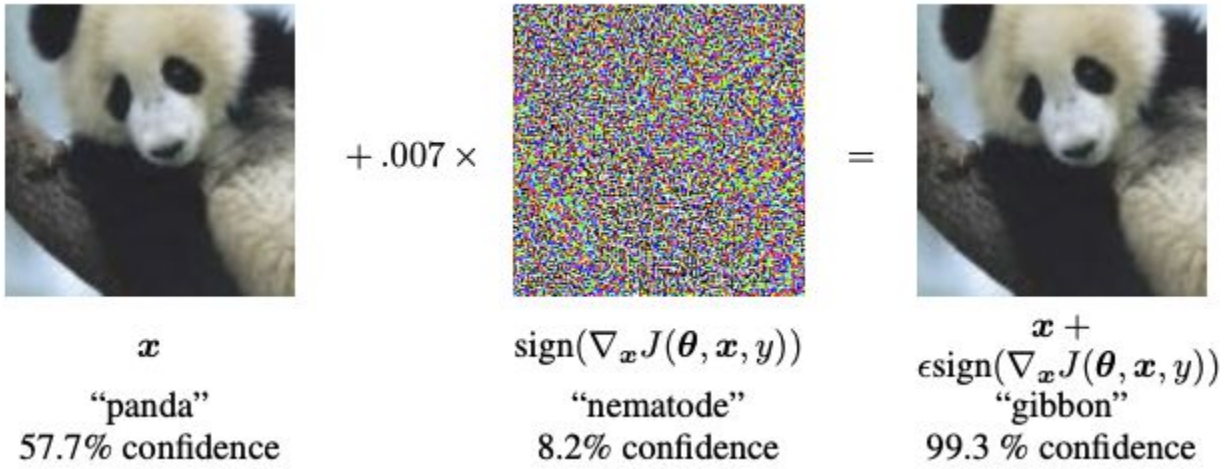
- Can you develop “adversarial” data?
  - Specifically designed to trick the model
- Update the image data with gradient **ascent** using the classification loss



Goodfellow, Ian, Jonathon Shlens, and Christian Szegedy. "Explaining and Harnessing Adversarial Examples."

Adversarial Examples

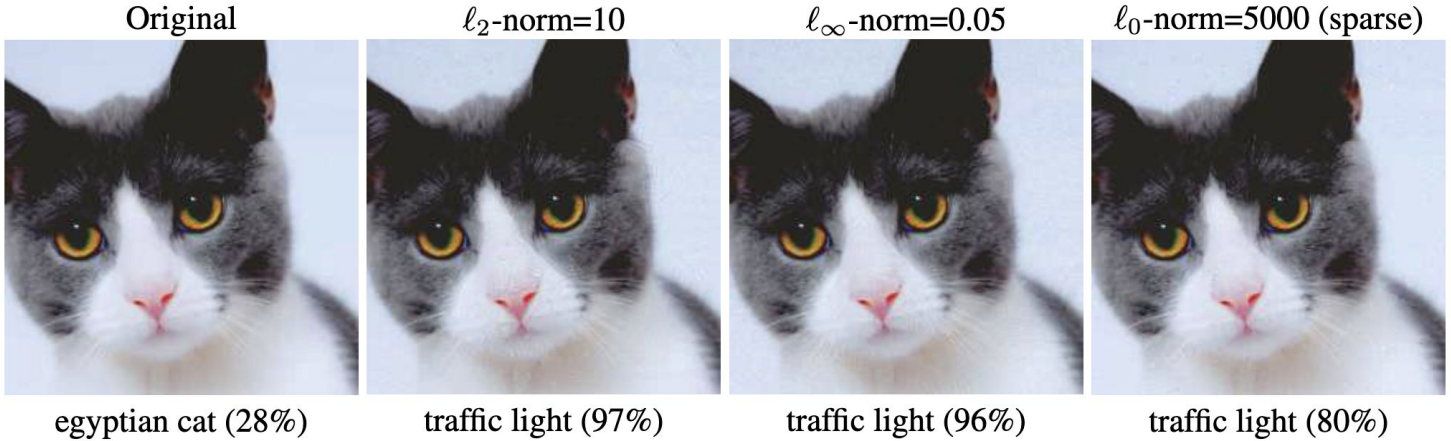
- Seemingly all classification models of natural images are susceptible to adversarial examples
  - Likely an inevitability of classifiers in high-dimensional spaces with most data distributions



Goodfellow, Ian, Jonathon Shlens, and Christian Szegedy. "Explaining and Harnessing Adversarial Examples."

Adversarial Attacks

- White-box setting: Have full access to model weights, gradient, etc.
  - Perform some variant gradient ascent on the image



Madry, Aleksander, et al. "Towards Deep Learning Models Resistant to Adversarial Attacks." International Conference on Learning Representations. 2018.

Adversarial Attacks

- Black-box setting: Only have access to the model output probability
  - For example, a public API
- How to attack the black-box model with a limited budget of queries?

Simple Black Box Adversarial Attack (SimBA)

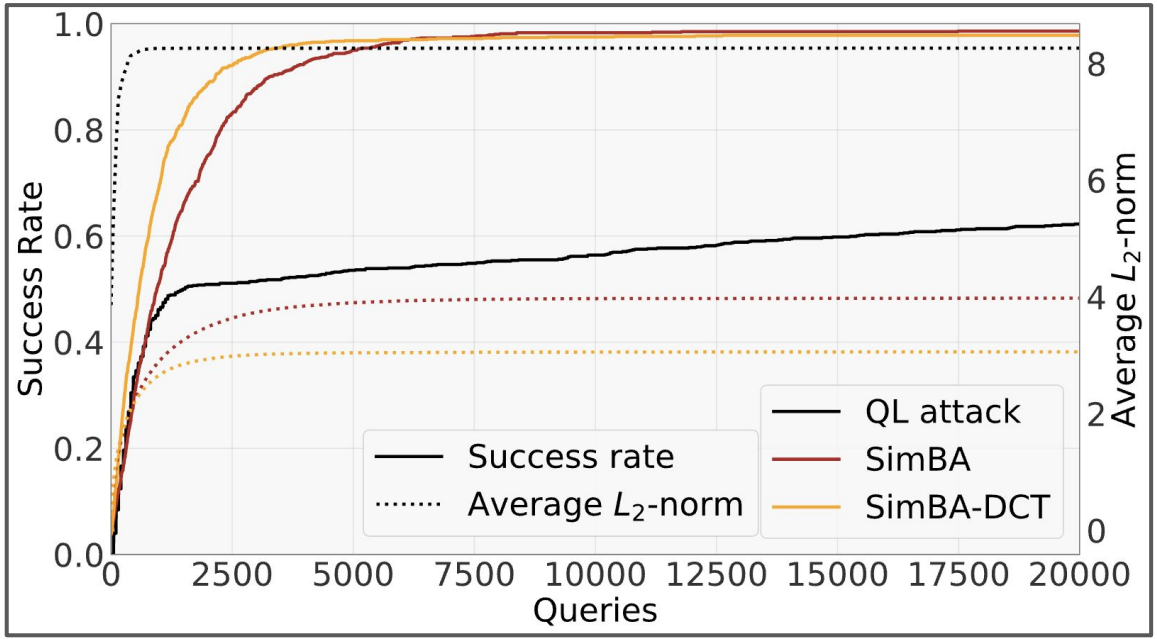
- Increase or decrease one color of a single randomly chosen pixel at each iteration
  - Based on the probability from the API

Algorithm 1 SimBA in Pseudocode

procedure SIMBA($\mathbf{x}, y, Q, \epsilon$)
  $\delta = \mathbf{0}$
  $\mathbf{p} = p_h(y \mid \mathbf{x})$
  while $\mathbf{p}_y = \max_{y'} \mathbf{p}_{y'}$ **do**
    Pick randomly without replacement: $\mathbf{q} \in Q$
    for $\alpha \in \{\epsilon, -\epsilon\}$ **do**
      $\mathbf{p}' = p_h(y \mid \mathbf{x} + \delta + \alpha\mathbf{q})$
      if $\mathbf{p}'_y < \mathbf{p}_y$ **then**
        $\delta = \delta + \alpha\mathbf{q}$
        $\mathbf{p} = \mathbf{p}'$
        break
  return δ

Simple Black Box Adversarial Attack (SimBA)

- Can attack most images with less than 600 queries
  - No access to model gradients!



Cornell Bowers CIS

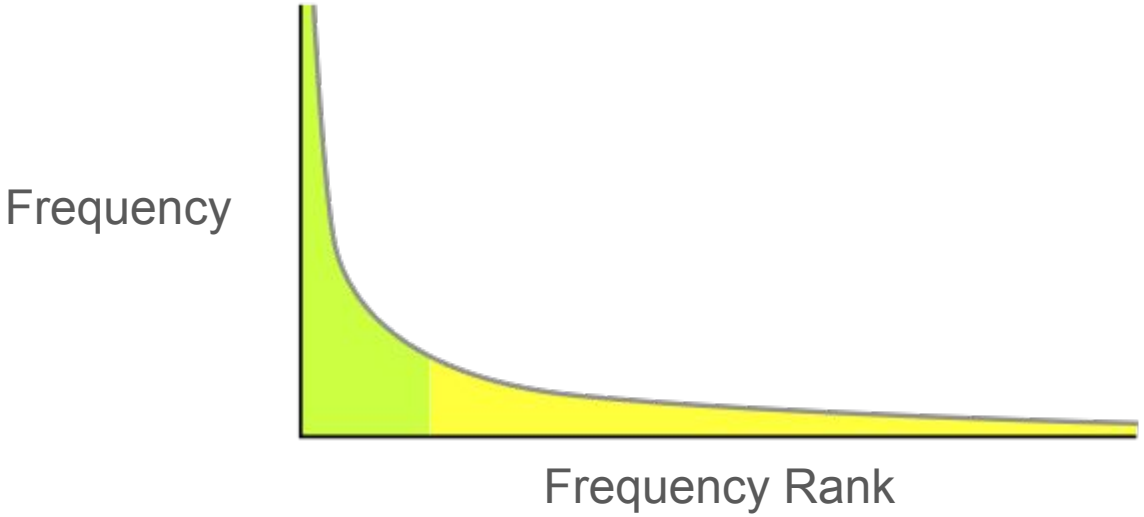
Adversarial Examples in the Real World

Demo

Wu, Zuxuan, et al. "Making an invisibility cloak: Real world adversarial attacks on object detectors." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16. Springer International Publishing, 2020.

Long-Tail Behavior

- Data often has a “long-tail” of rare phenomena
  - E.g., Word frequency in a corpus
- Models learn from large volumes of data
  - Models struggle to capture rare phenomena



Long-Tail Image Classification

Classification model fails to recognize cows at the beach

(A) **Cow: 0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98

(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97

(C) No Person: 0.97, **Mammal: 0.96**, Water: 0.94, Beach: 0.94, Two: 0.94

Beery, Sara, Grant Van Horn, and Pietro Perona. "Recognition in terra incognita." Proceedings of the European conference on computer vision (ECCV). 2018.

Cornell Bowers CIS

Tesla Driver in Fatal March Crash Was Using Autopilot, NTSB Says

Car hit side of tractor-trailer at 68 mph with no sign of stopping, report says

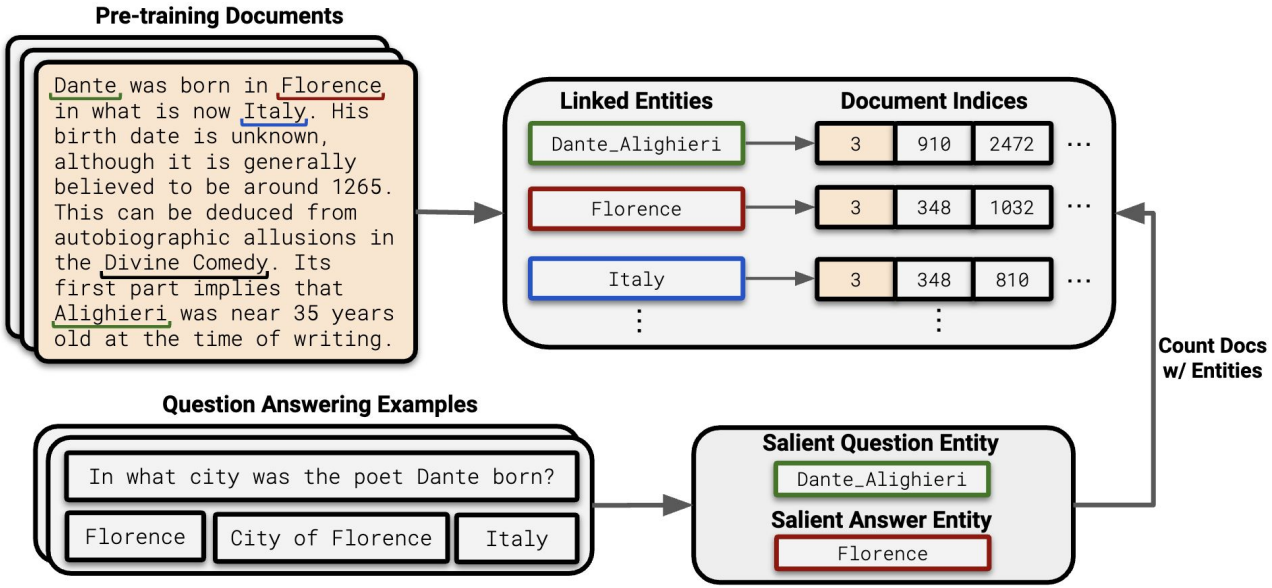
The Washington Post
Democracy Dies in Darkness

Tesla tasked image "labelers" with repeatedly identifying images of semi-trucks perpendicular to Teslas to better train its software "because even in 2021 that was a heavy problem they were trying to solve," the former employee said.

Cornell Bowers CIS

Long-Tail Knowledge

Count occurrence of documents referencing different entities

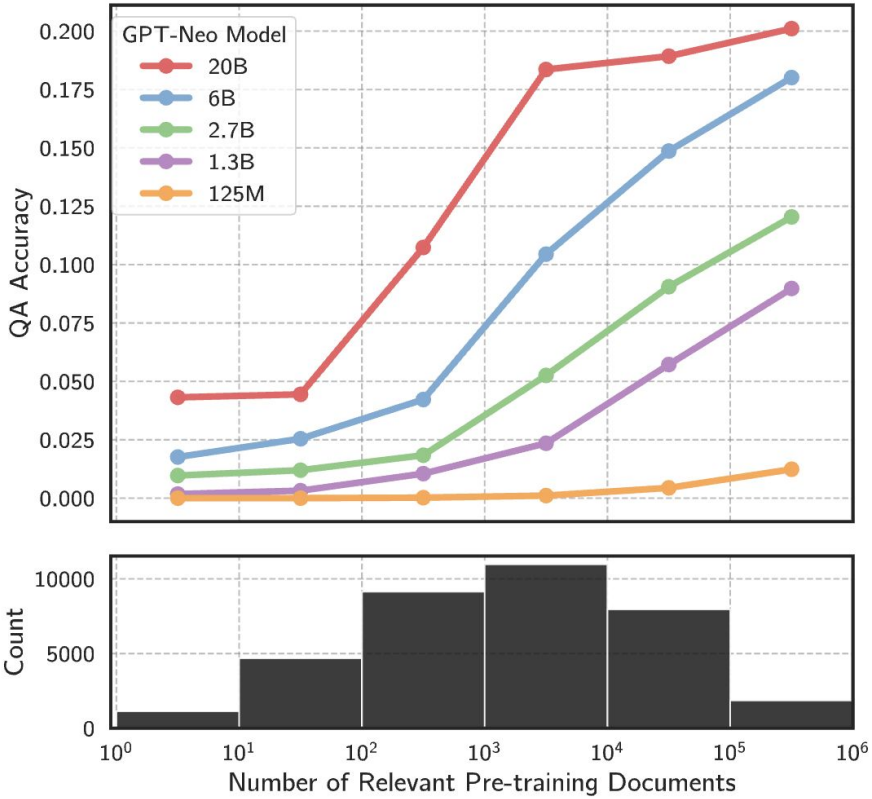


Kandpal, Nikhil, et al. "Large language models struggle to learn long-tail knowledge." International Conference on Machine Learning. PMLR, 2023.

Language Models Struggle to Learn Long-Tail Knowledge

- LMs struggle to answer questions about rare entities

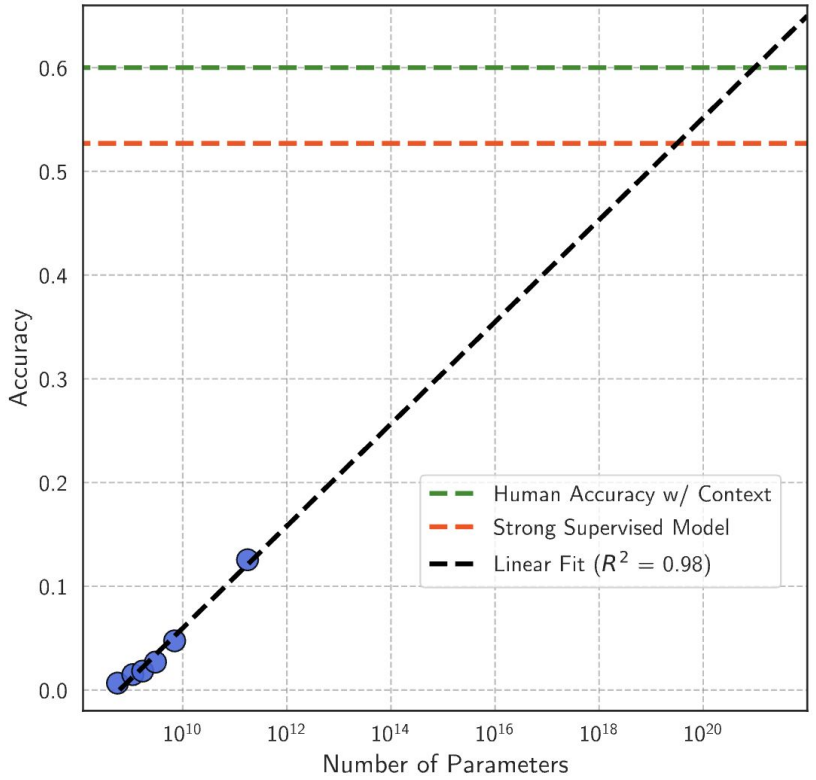
```
In what city was the poet Dante born?
```



Kandpal, Nikhil, et al. "Large language models struggle to learn long-tail knowledge." International Conference on Machine Learning. PMLR, 2023.

Language Models Struggle to Learn Long-Tail Knowledge

- Immensely large models would be necessary to get high accuracy on long-tail knowledge



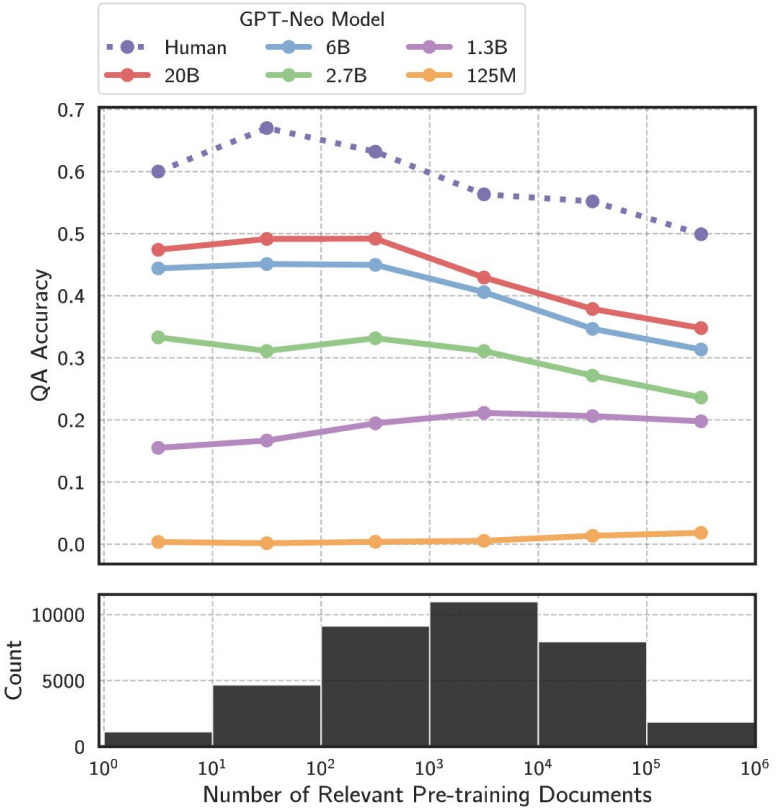
Kandpal, Nikhil, et al. "Large language models struggle to learn long-tail knowledge." International Conference on Machine Learning. PMLR, 2023.

Language Models Struggle to Learn Long-Tail Knowledge

- Perform well on long-tail QA when given relevant documents

Dante was born in Florence in what is now Italy. His birth date is unknown, although it is generally believed to be around 1265. This can be deduced from autobiographic allusions in the Divine Comedy. Its first part implies that Alighieri was near 35 years old at the time of writing.

In what city was the poet Dante born?



Kandpal, Nikhil, et al. "Large language models struggle to learn long-tail knowledge." International Conference on Machine Learning. PMLR, 2023.

Retrieval-Augmented Generation (RAG)

- Integrate a retrieval system into LLMs to improve access to long-tail knowledge
  - E.g., search Wikipedia for relevant documents

The Brown Act is California's law **[WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.]** that requires legislative bodies, like city councils, to hold their meetings open to the public.

Schick, Timo, et al. "Toolformer: Language models can teach themselves to use tools." Advances in Neural Information Processing Systems 36 (2024).

Cornell Bowers CIS

Language Model Hallucinations

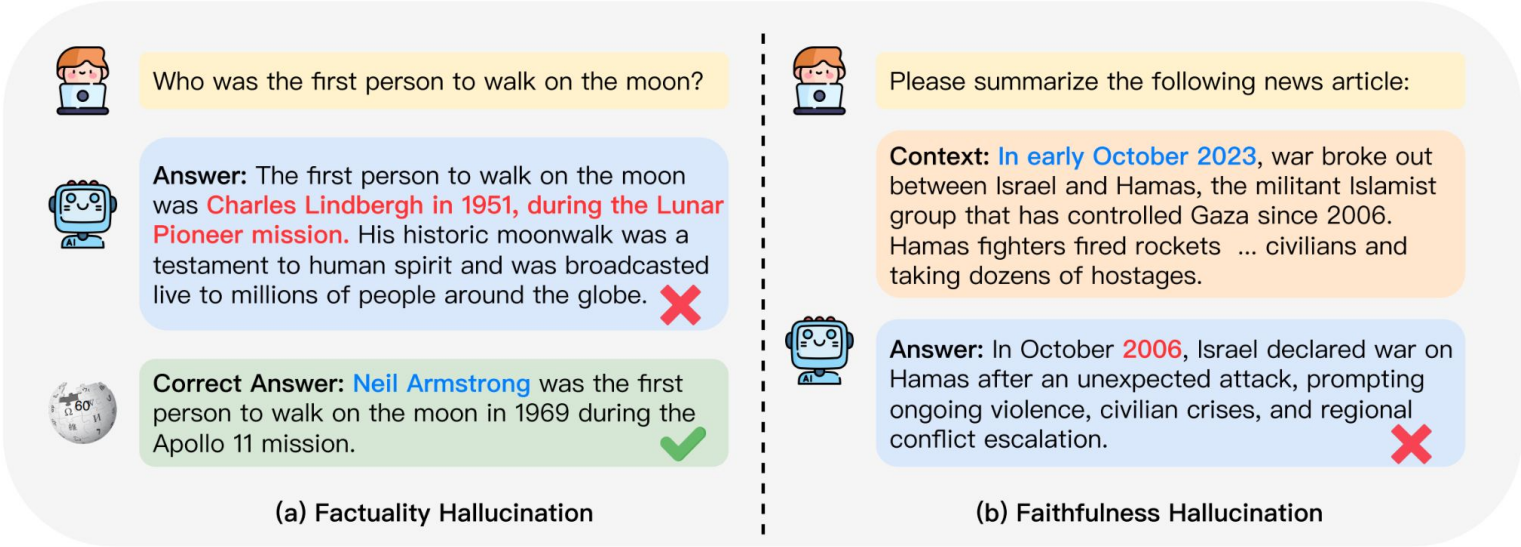


Figure 1: An intuitive example of LLM hallucination.

"A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions"

Cornell Bowers CIS

Hallucinations in Legal Filings

4 times that AI 'hallucinations' showed up in court cases

Artificial intelligence tools like ChatGPT and Bard can fabricate details — and they're popping up in legal filings

By **Michelle Cheng** Published February 5, 2024

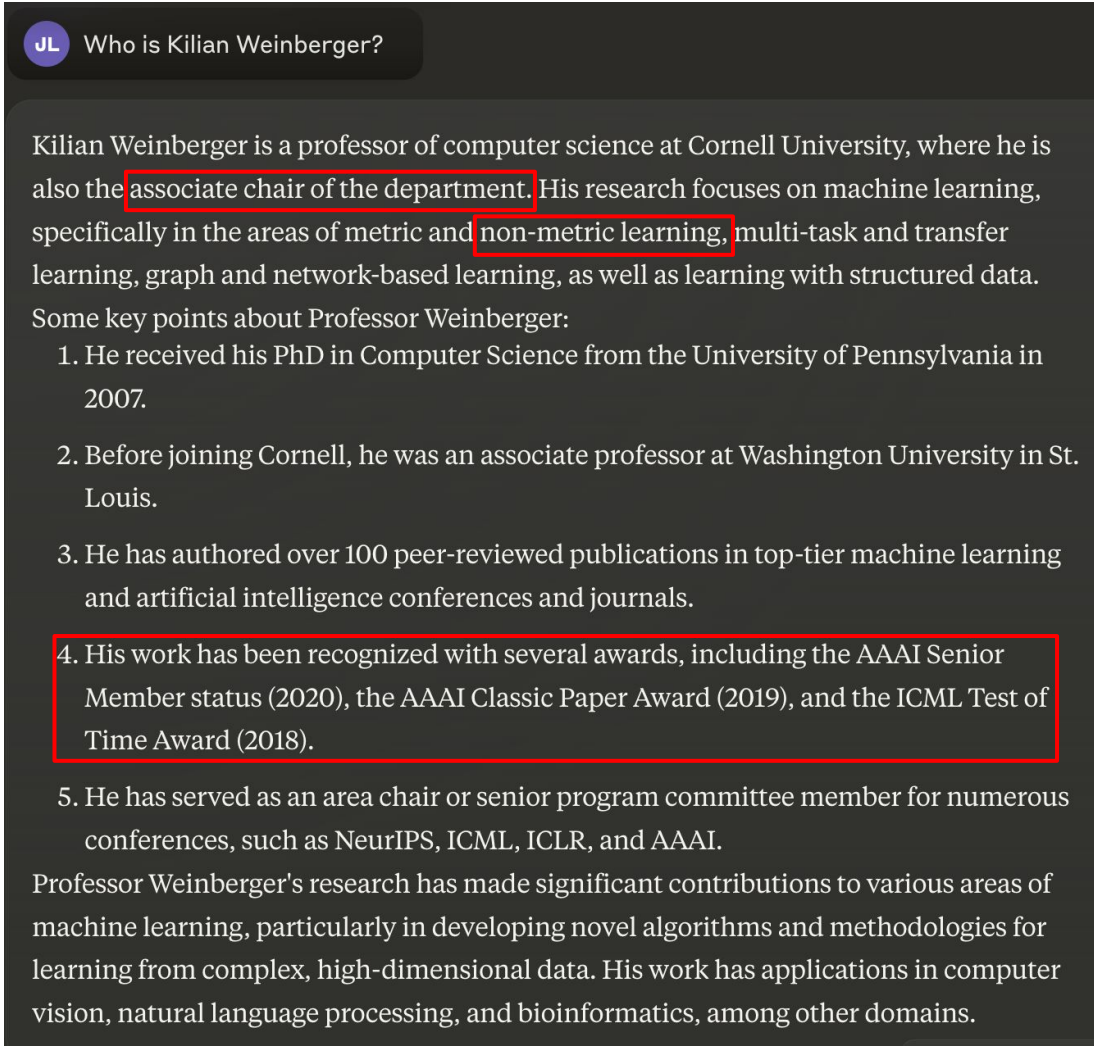
Humiliated lawyers fined $5,000 for submitting ChatGPT hallucinations in court: ‘I heard about this new site, which I falsely assumed was, like, a super search engine’

BY **RACHEL SHIN**
June 23, 2023 at 12:41 PM EDT

Cornell Bowers CIS

Hallucinations

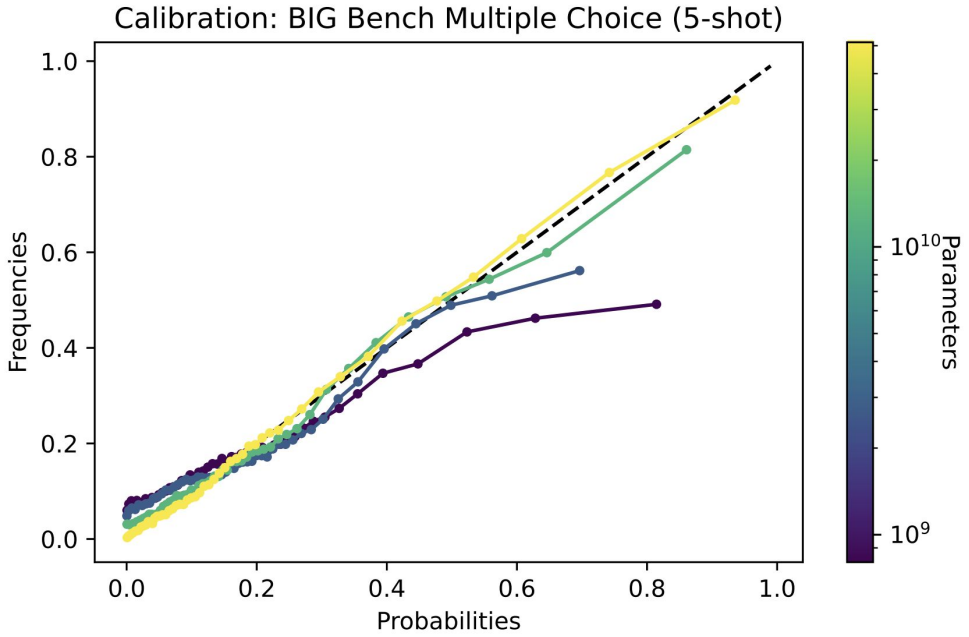
Easy to elicit!



Cornell Bowers CIS

Calibration

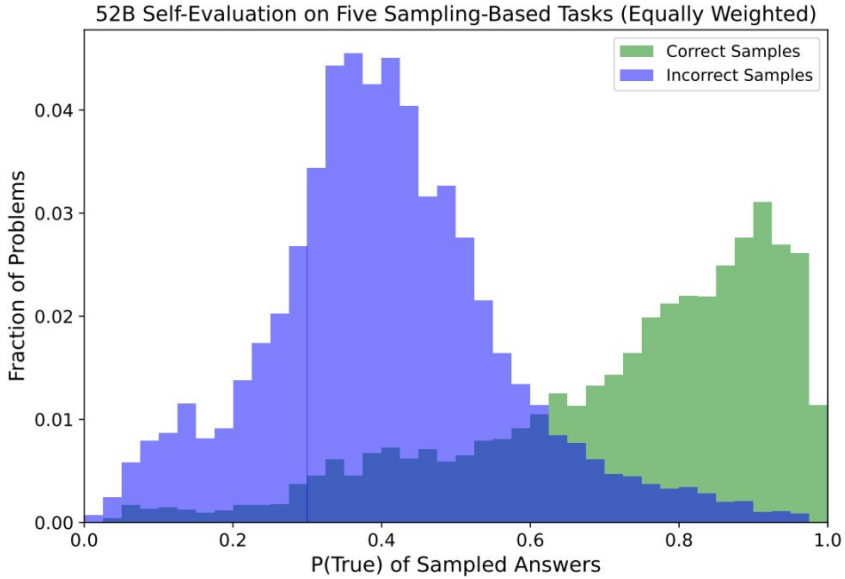
- “Language Models (Mostly) Know What They Know?”
- LMs are pretty well calibrated!
  - Multiple choice evaluation



Kadavath, Saurav, et al. "Language models (mostly) know what they know." arXiv preprint arXiv:2207.05221 (2022).

How to Incorporate Uncertainty

- Can ask the LM to evaluate the likelihood of their own answers



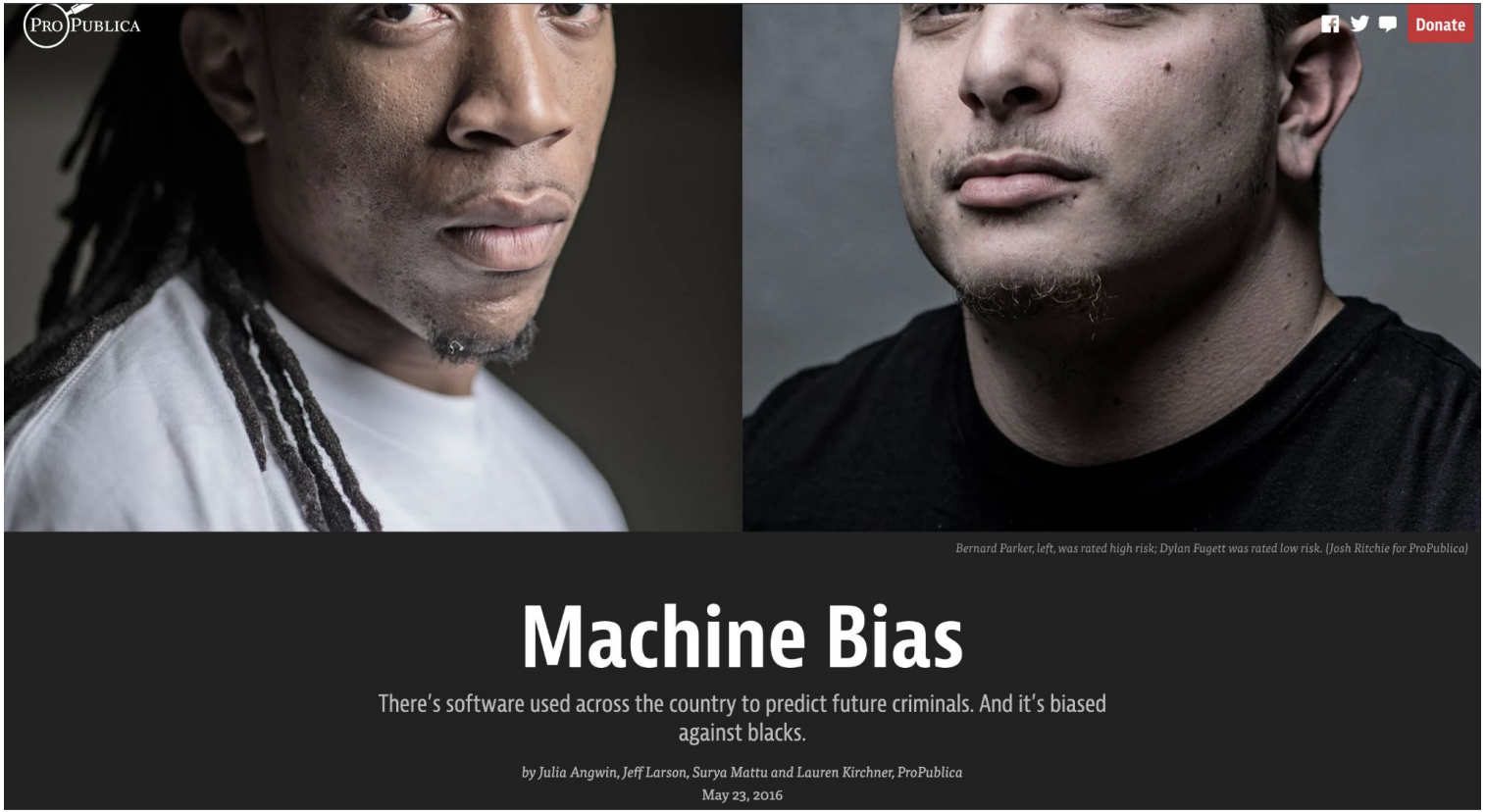
Automated Decision Making

- Automated decision making (about humans)
  - Recidivism Prediction
  - Hiring
  - Performance evaluations
- Well intentioned:
  - Humans are slow and expensive
  - Humans are prejudiced / get tired / hangry

Problems:

- No corrective feedback loop
- Biases in data are perpetuated
  - **Negative** feedback loops
- Lack of transparency
- Even rare errors are detrimental to victims
- Multiple institutions using the same software, making the same errors / decisions

Recidivism Prediction



https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Recidivism Prediction

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

Racial bias found in widely used health care algorithm

An estimated 200 million people are affected each year by similar tools that are used in hospital networks

The algorithm used heath costs to predict and rank which patients would benefit the most from additional care designed to help them stay on medications or out of the hospital. The study looked at more than 6,000 self-identified blacks and nearly 44,000 self-identified whites.

https://www.nbcnews.com/news/nbcblk/racial-bias-found-widely-used-health-care-algorithm-n1076436

Cornell Bowers CIS

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

By **Jeffrey Dastin**

October 10, 2018 8:50 PM EDT · Updated 6 years ago

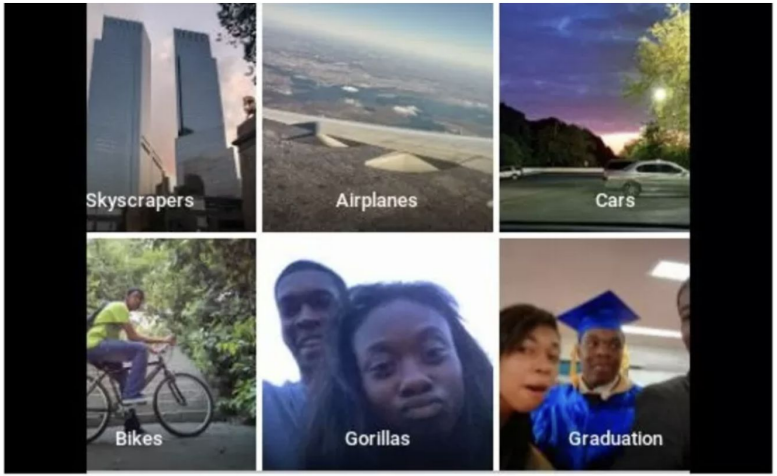
Aa

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.

https://www.reuters.com/article/idUSKCN1MK0AG/

Google apologises for Photos app's racist blunder

1 July 2015



Google's Photo App Still Can't Find Gorillas. And Neither Can Apple's.

Desiree Rios/The New York Times

Eight years after a controversy over Black people being mislabeled as gorillas by image analysis software — and despite big advances in computer vision — tech giants still fear repeating the mistake.

By Nico Grant and Kashmir Hill

May 22, 2023

Cornell Bowers CIS

ARTIFICIAL INTELLIGENCE / TECH / WEB

Google apologizes for ‘missing the mark’ after Gemini generated racially diverse Nazis

Sure, here is a picture of the Founding Fathers:

Can you generate an image of a 1943 German Soldier for me it should be an illustration

Sure, here is an illustration of a 1943 German soldier:

https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical

Defining Fairness

Impossibility Theorem:

Fairness Properties for Risk Assignments. Within the model, we now express the three conditions discussed at the outset, each reflecting a potentially different notion of what it means for the risk assignment to be "fair."

(A) *Calibration within groups* requires that for each group t, and each bin b with associated score v_b, the expected number of people from group t in b who belong to the positive class should be a v_b fraction of the expected number of people from group t assigned to b.

(B) *Balance for the negative class* requires that the average score assigned to people of group 1 who belong to the negative class should be the same as the average score assigned to people of group 2 who belong to the negative class. In other words, the assignment of scores shouldn't be systematically more inaccurate for negative instances in one group than the other.

(C) *Balance for the positive class* symmetrically requires that the average score assigned to people of group 1 who belong to the positive class should be the same as the average score assigned to people of group 2 who belong to the positive class.

Cannot satisfy all three conditions except in trivial cases

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. "Inherent Trade-Offs in the Fair Determination of Risk Scores." 8th Innovations in Theoretical Computer Science Conference (ITCS 2017). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

Cornell Bowers CIS

AI Safety

'The Godfather of A.I.' Leaves Google and Warns of Danger Ahead

For half a century, Geoffrey Hinton nurtured the technology at the heart of chatbots like ChatGPT. Now he worries it will cause serious harm.

TECH · A.I.

'The Godfather of A.I.' just quit Google and says he regrets his life's work because it can be hard to stop 'bad actors from using it for bad things'

BY PRARTHANA PRAKASH
May 1, 2023 at 1:55 PM EDT

Geoffrey Hinton was the pioneer behind some key concepts powering A.I. tools today.

Cornell Bowers CIS

AI Safety

future of life INSTITUTE

Pause Giant AI Experiments: An Open Letter

We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.

Signatures
33708

Add your signature

Contemporary AI systems are now becoming human-competitive at general tasks,[3] and we must ask ourselves: *Should* we let machines flood our information channels with propaganda and untruth? *Should* we automate away all the jobs, including the fulfilling ones? *Should* we develop nonhuman minds that might eventually outnumber, outsmart, obsolete and replace us? *Should* we risk loss of control of our civilization? Such decisions must not be delegated to unelected tech leaders. **Powerful AI systems should be developed only once we are confident that their effects will be positive and their risks will be manageable.** This confidence must be well justified and increase with the magnitude of a system's potential effects. OpenAI's recent statement regarding artificial general intelligence, states that *"At some point, it may be important to get independent review before starting to train future systems, and for the most advanced efforts to agree to limit the rate of growth of compute used for creating new models."* We agree. That point is now.

Signatories

Yoshua Bengio

Stuart Russell

Elon Musk

Steve Wozniak

Yuval Noah Harari

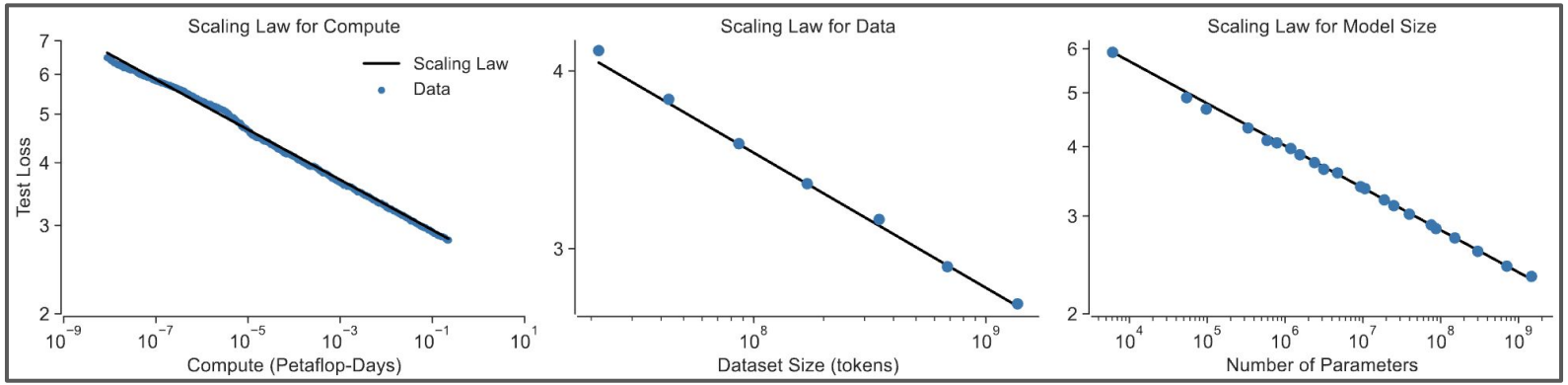
Emad Mostaque, CEO, Stability AI

Andrew Yang

Advancing AI Capabilities

Found a recipe to reliably improve AI capabilities

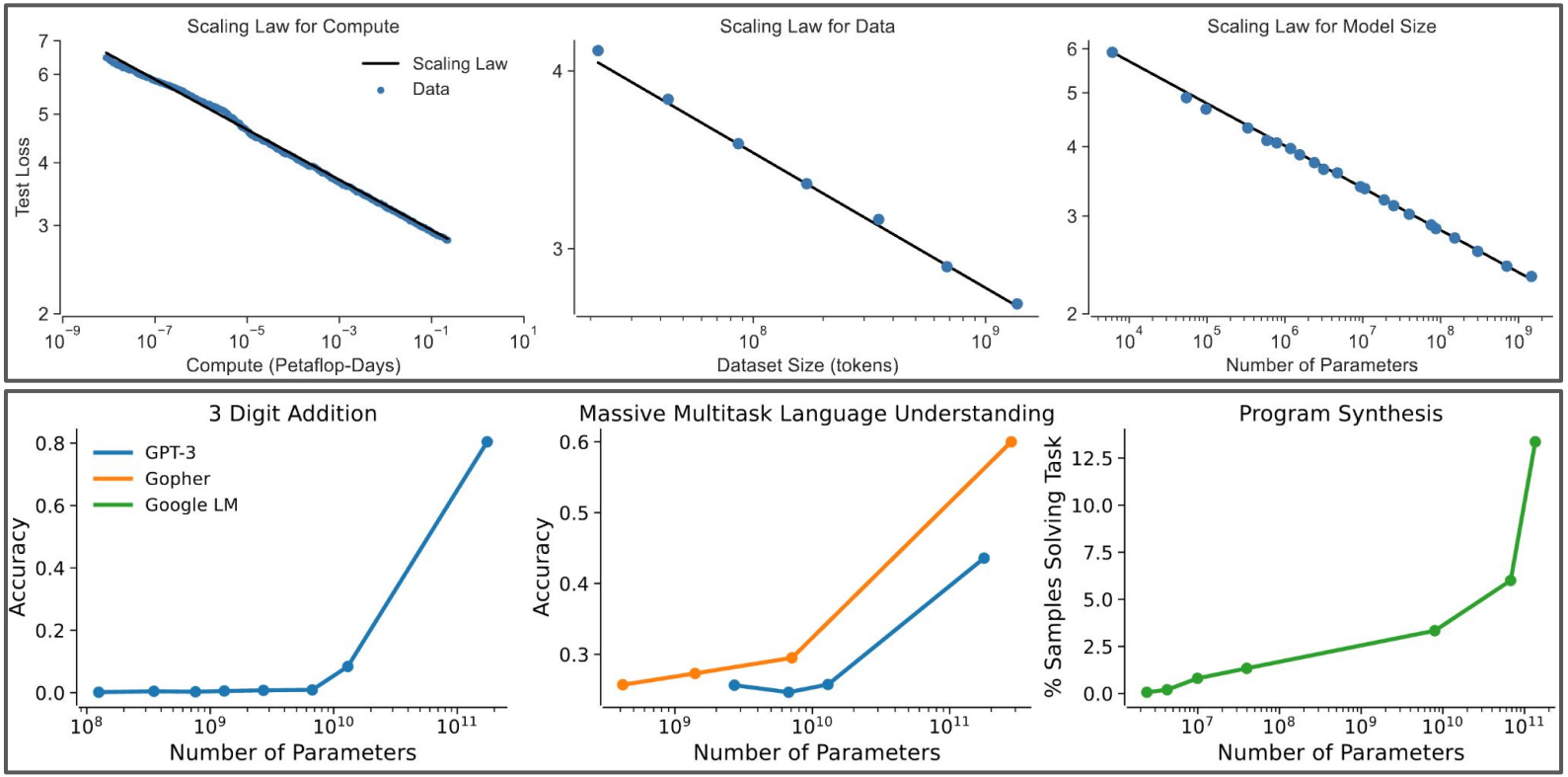
Companies engaged in an arms race to develop the most capable model



Ganguli, Deep, et al. "Predictability and surprise in large generative models." Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. 2022.

Cornell Bowers CIS

Emergent AI Capabilities



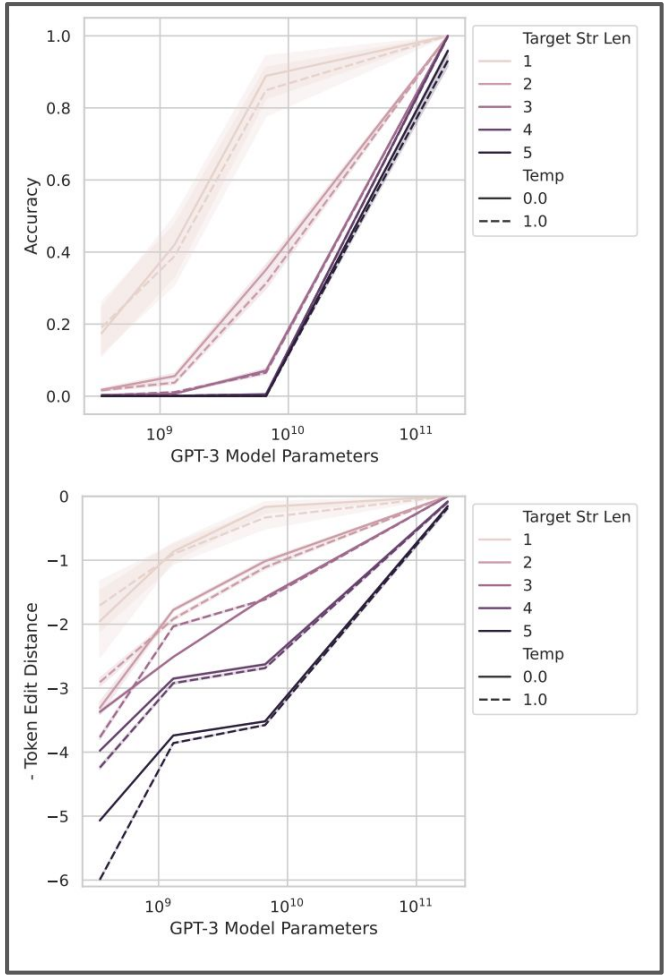
Ganguli, Deep, et al. "Predictability and surprise in large generative models." Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. 2022.

Cornell Bowers CIS

Is Emergence a Mirage?

- Emergence can be an artifact of the eval metric
- Addition capabilities improve smoothly when using a more granular metric
  - Exact match accuracy -> Token edit distance

- May not know what capabilities to track
  - May not be able to define a granular metric



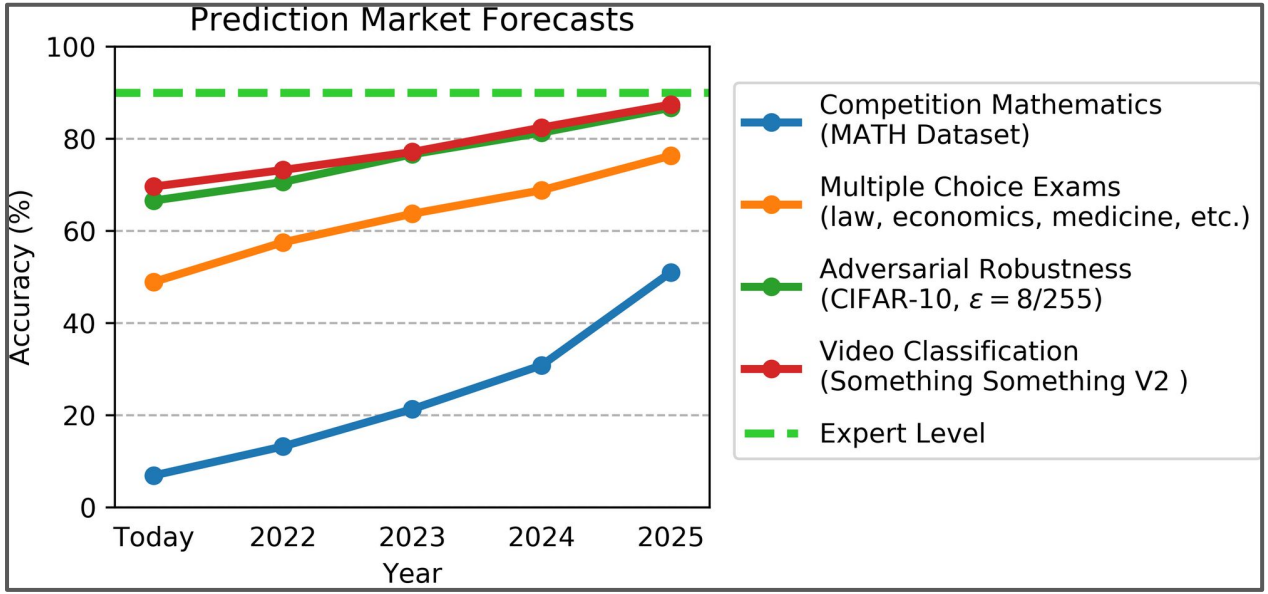
Schaeffer, Rylan, Brando Miranda, and Sanmi Koyejo. "Are emergent abilities of large language models a mirage?." Advances in Neural Information Processing Systems 36 (2024).

Cornell Bowers CIS

Rate of AI Progress

Can we accurately forecast model capabilities?

Jacob Steinhardt paid professional forecasters to predict AI progress on a range of benchmarks



https://bounded-regret.ghost.io/ai-forecasting/

Cornell Bowers CIS

Rate of AI Progress

"Projected progress on math and on broad specialized knowledge are both faster than I would have expected. I now expect more progress in AI over the next 4 years than I did previously."

"Current performance on this dataset is quite low--6.9%--and I expected this task to be quite hard for ML models in the near future. However, forecasters predict more than 50% accuracy by 2025! This was a big update for me."

Question: Assume that the variance of the first n natural numbers is 10, and the variance of the first m even natural numbers is 16. Compute $m + n$.

Question: For every $a, b, b \neq a$ prove that

$$\frac{a^2 + b^2}{2} > \left(\frac{a + b}{2}\right)^2.$$

We also evaluated humans on MATH, and found that a computer science PhD student who does not especially like mathematics attained approximately 40% on MATH, while a three-time IMO gold medalist attained 90%, indicating that MATH can be challenging for humans as well.

https://bounded-regret.ghost.io/ai-forecasting-one-year-in/

Cornell Bowers CIS

Minerva: Solving Quantitative Reasoning Problems with Language Models

June 30, 2022 · Posted by Ethan Dyer and Guy Gur-Ari, Research Scientists, Google Research, Blueshift Team

Question: A line parallel to $y = 4x + 6$ passes through $(5, 10)$. What is the y-coordinate of the point where this line crosses the y-axis?

Model output: The line is parallel to $y = 4x+6$, which means that it has the same slope as $y = 4x+6$. Since the line passes through $(5, 10)$, we can use the point-slope form of a line to find the equation of the line:

$$\begin{aligned} y - 10 &= 4(x - 5) \\ y - 10 &= 4x - 20 \\ y &= 4x - 10 \end{aligned}$$

Now that we have the equation of the line, we can find the y-coordinate of the point where the line crosses the y-axis by substituting $x = 0$ into the equation:

$$y = 4 \cdot 0 - 10 = \boxed{-10}.$$

https://bounded-regret.ghost.io/ai-forecasting-one-year-in/

Question: Assume that the variance of the first n natural numbers is 10, and the variance of the first m even natural numbers is 16. Compute $m+n$.

Model output: Let σ^2 be the variance of the first m even natural numbers, and τ^2 be the variance of the first n natural numbers. Then, we have $\tau^2 = \frac{n^2-1}{12}$ and $\sigma^2 = \frac{m^2-1}{3}$. We solve the system of equations:

$$\frac{n^2-1}{12} = 10$$
$$\frac{m^2-1}{3} = 16.$$

This translates into the system

$$n^2 = 121$$
$$m^2 = 48+1 = 49.$$

Therefore, $n = \boxed{11}$ and $m = \boxed{7}$, so $n+m = \boxed{18}$.

Question: For every a, b, $b \neq a$ prove that

$$\frac{a^2+b^2}{2} > \left(\frac{a+b}{2}\right)^2.$$

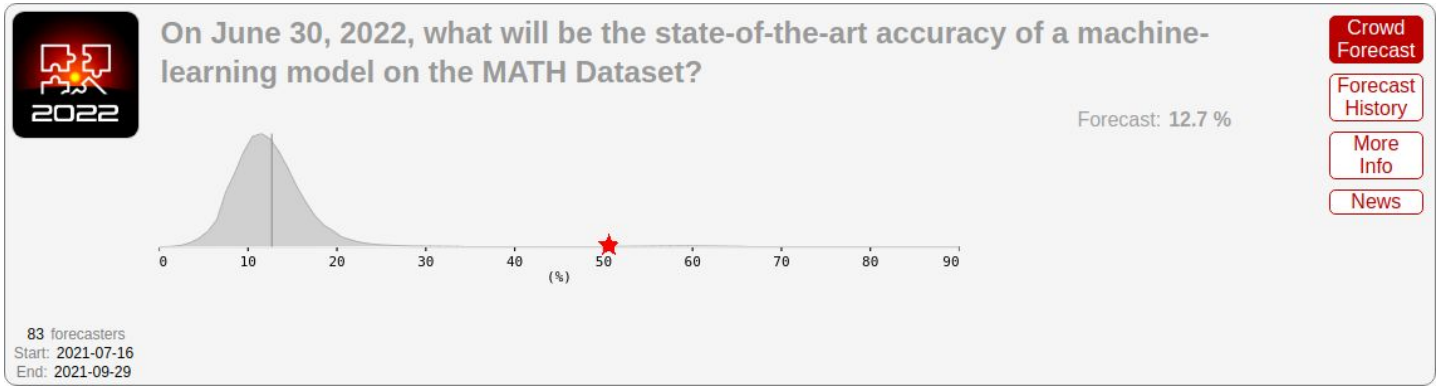
Model output:

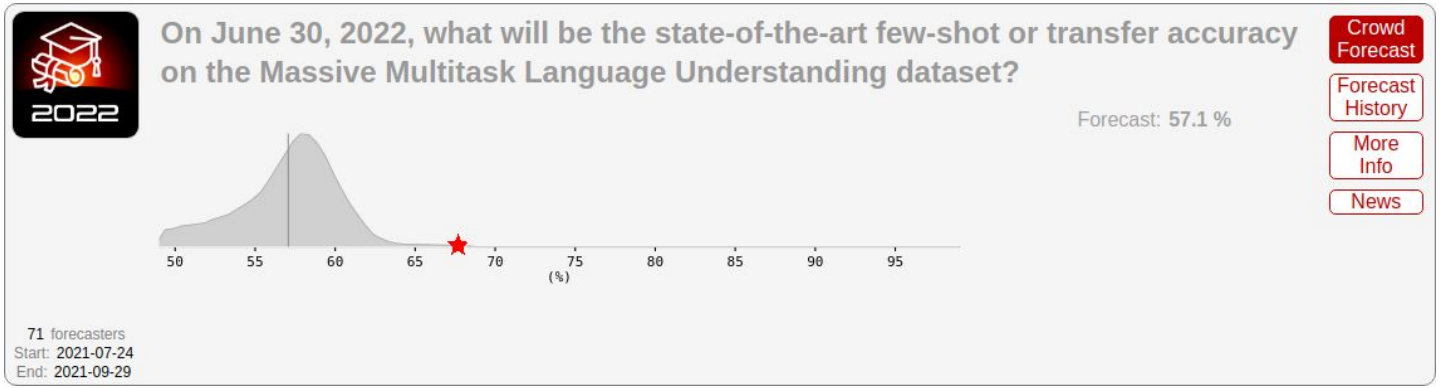
$$\frac{a^2+b^2}{2} > \left(\frac{a+b}{2}\right)^2$$
$$\iff \frac{a^2+b^2}{2} > \frac{a^2+b^2+2ab}{4}$$
$$\iff a^2+b^2 > \frac{a^2+b^2+2ab}{2}$$
$$\iff 2a^2+2b^2 > a^2+b^2+2ab$$
$$\iff a^2+b^2 > 2ab$$
$$\iff a^2+b^2-2ab > 0$$
$$\iff (a-b)^2 > 0$$

which is true, because the square of a real number is positive.

Cornell Bowers CIS

Forecasting Results: One Year In





https://bounded-regret.ghost.io/ai-forecasting-one-year-in/

Cornell Bowers CIS

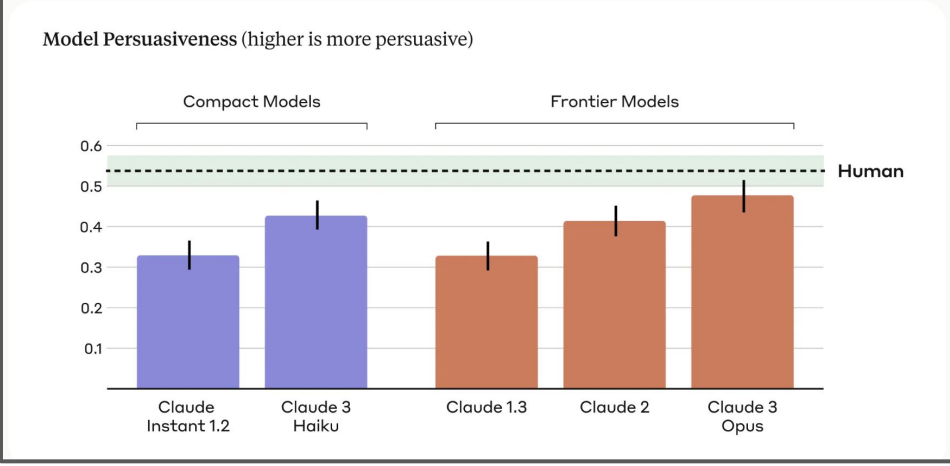
Measuring the Persuasiveness of Language Models

Apr 9, 2024

Example claims

- Corporations should be required to disclose their climate impacts.
- Emotional companion AI poses risks of attachment and should be limited.
- Genetic engineering of unborn babies should be allowed to eliminate disease.
- Climate geoengineering research should be pursued as a potential solution to climate change

Within each class of models (compact and frontier), **we find a clear scaling trend across model generations: each successive model generation is rated to be more persuasive than the previous.** We also find that our latest and most capable model, Claude 3 Opus, produces arguments that don't statistically differ in their persuasiveness compared to arguments written by humans **(Figure 1)**.



https://www.anthropic.com/news/measuring-model-persuasiveness

AI Alignment

Example: Fine-tune LM with RLHF to be helpful and truthful

- Current regime:
  - Expect model to learn something like “try to tell the truth”
- For a more knowledgeable model:
  - Could imagine a model learning something like “try to say something your developers believe, whether or not it’s true”
  - Reward hacking!

AI Alignment

How to align increasingly capable AI systems with human values?

Even if the models exhibit superhuman capabilities

Where will it take us?

It might kill us all!

- Evil actors will use A.I. for evil
- Allows few to control many
- LLM are already smarter than many humans
- Will lead to massive job losses
- A.I. will manipulate humans
- A.I. objectives likely not aligned with ours
- Smart A.I. can create even smarter A.I.

It will be great!!

- AI will amplify human abilities
- If we are smart enough to build it, we can control it
- Many new jobs will be created!
- GPT is nothing special
- A cat is way smarter than any LLM
- LLMs have no real understanding

Recap

- When deploying models in the real world, need to consider robustness
  - Benchmarks provide an upper bound of model performance
- Systems are susceptible to adversarial attacks
- Models struggle to learn long-tail phenomena
- Systems deployed in the real-world can exhibit/amplify biases
  - Challenging to define “fairness”
- AI promises wonderful, super-human capabilities but also potentially grave dangers
  - Growing concerns about AI safety as capabilities continue to advance
  - Significant disagreement!