Natural Language Processing

Berkeley

N L P

Large Language Models

RLHF: Reinforcement Learning from Human Feedback

PROMPT	*Explain the moon landing to a 6 year old in a few sentences.*
COMPLETION	GPT-3

```
Explain the theory of gravity to a 6 year old.

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

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.
```

Ouyang et al. 2022

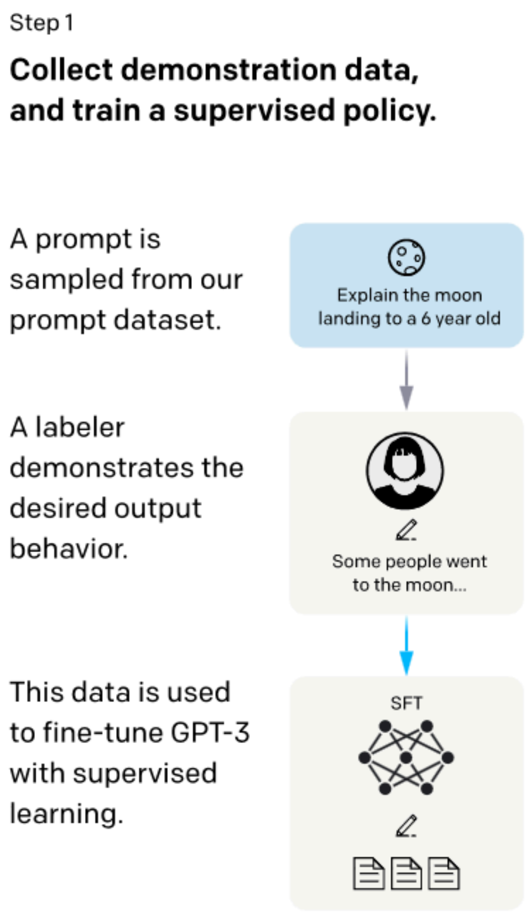
RLHF: Reinforcement Learning from Human Feedback

Main idea:
augment
training by
getting labels
for new
generations
using RL

Ouyang et al. 2022

RLHF: Reinforcement Learning from Human Feedback

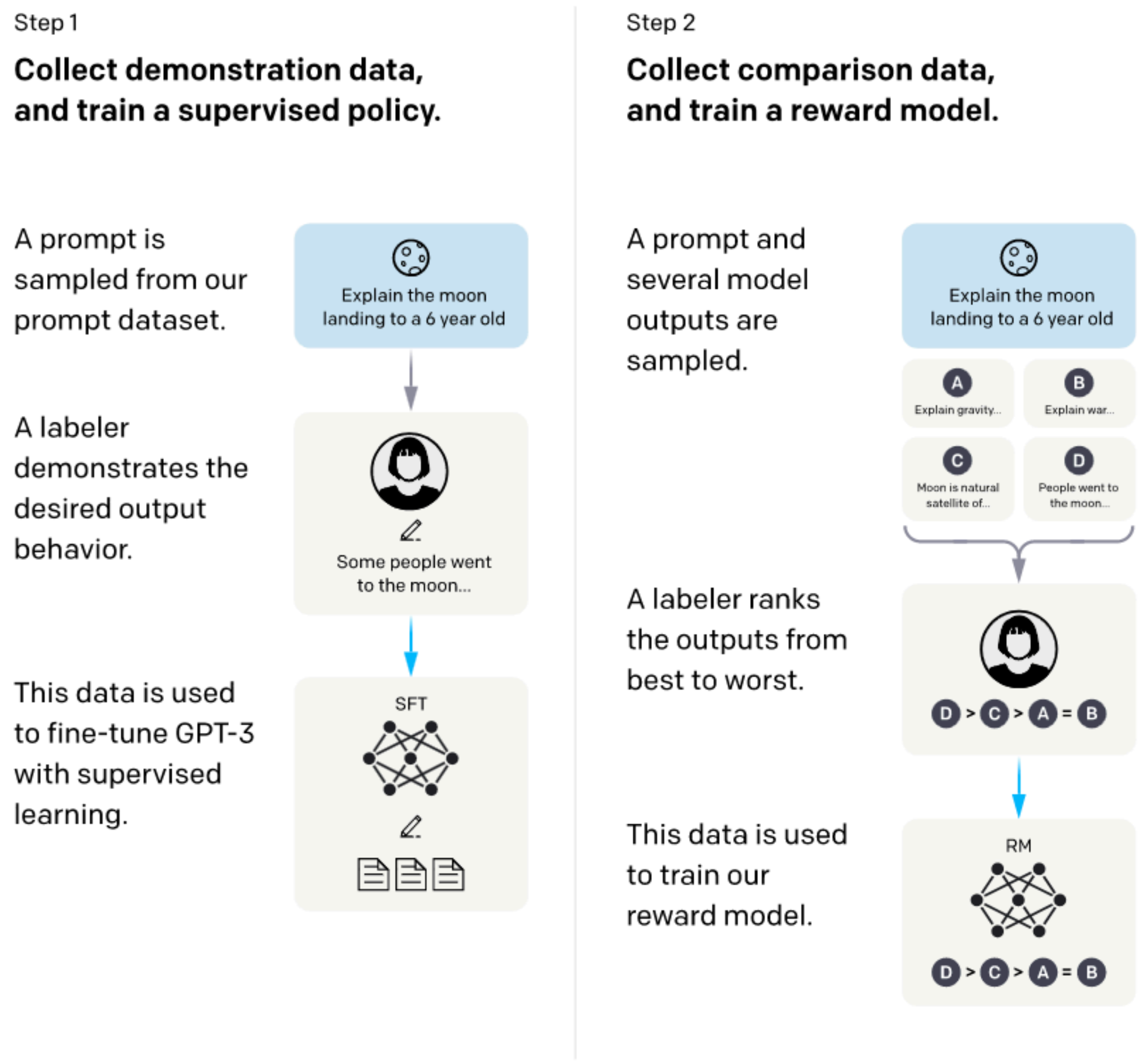
Main idea: augment training by getting labels for new generations using RL



Ouyang et al. 2022

RLHF: Reinforcement Learning from Human Feedback

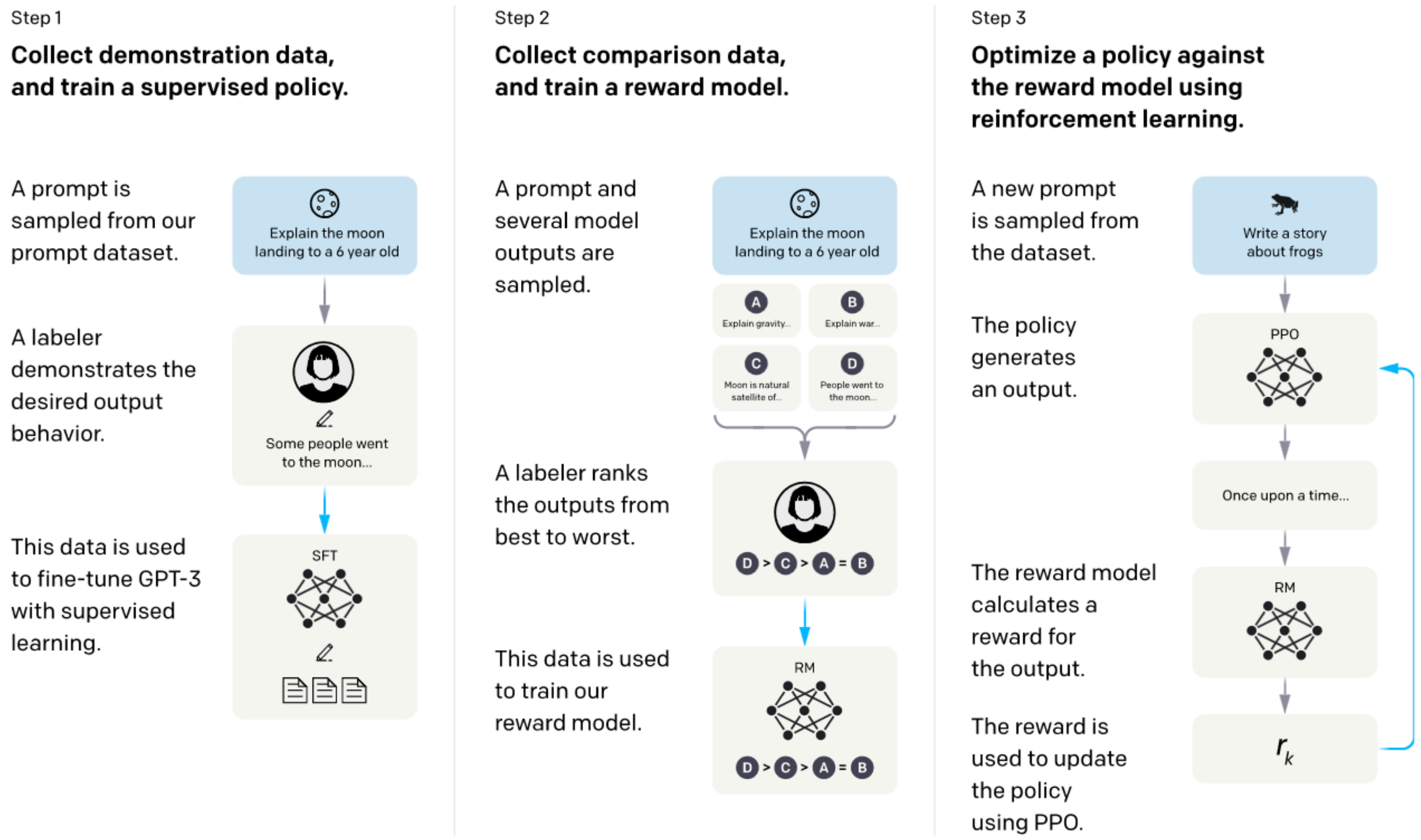
Main idea: augment training by getting labels for new generations using RL



Ouyang et al. 2022

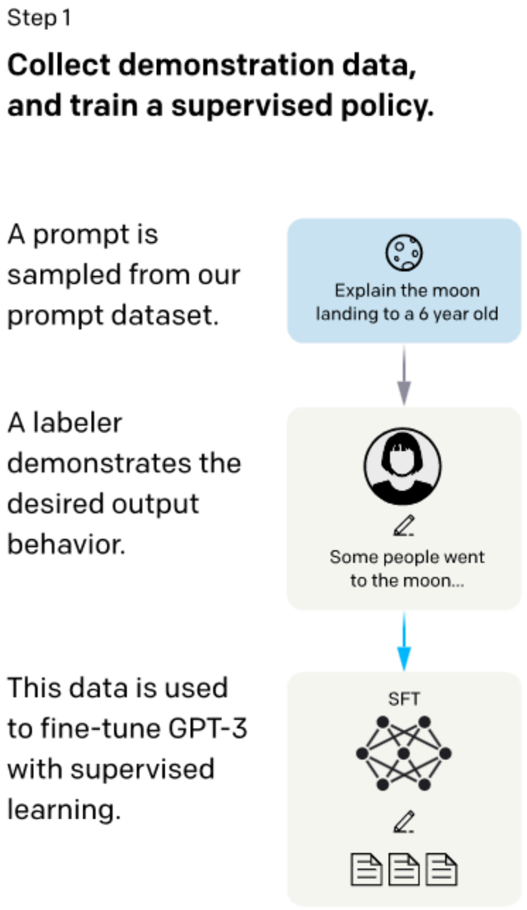
RLHF: Reinforcement Learning from Human Feedback

Main idea: augment training by getting labels for new generations using RL



Ouyang et al. 2022

RLHF: Supervised Fine-Tuning



PROMPT	*Explain the moon landing to a 6 year old in a few sentences.*
COMPLETION	**Human** A giant rocket ship blasted off from Earth carrying astronauts to the moon. The astronauts landed their spaceship on the moon and walked around exploring the lunar surface. Then they returned safely back to Earth, bringing home moon rocks to show everyone.

$$p \sim \mathcal{D}_p$$

$$\bar{x} = \mathrm{HumanDemonstration}(p)$$

$$\mathcal{D}_d = \mathcal{D}_d \cup \{p\bar{x}\}$$

Initial θ is GPT-3's parameters.

$$\theta_{\mathrm{sup}} \approx \arg\max_{\theta} \mathbb{E}_{d \in \mathcal{D}_d} \log(\pi_\theta(d))$$

Ouyang et al. 2022

RLHF: Training the Reward Model

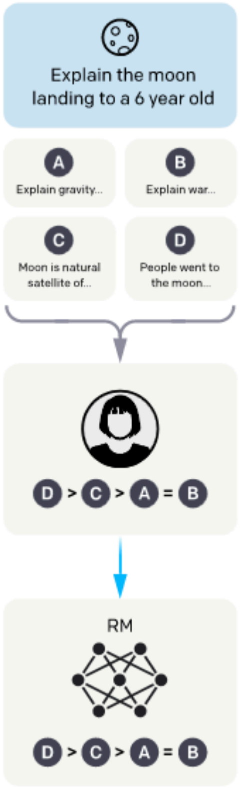
Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



$$p \sim \mathcal{D}_p$$

$$\tilde{\mathcal{X}} \sim \pi_{\theta_{\mathrm{sup}}}(\cdot \mid p)$$

Sample between 4 and 9 continuations per prompt.

$$\langle \tilde{x}_0, \ldots, \tilde{x}_N \rangle = \mathrm{HumanRanking}(p, \tilde{\mathcal{X}})$$

Some outputs might be rated equivalent.

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

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

Table 2: Illustrative prompts from our API prompt dataset. These are fictional examples inspired by real usage—see more examples in Appendix A.2.1.

Use-case	Prompt
Brainstorming	List five ideas for how to regain enthusiasm for my career
Generation	Write a short story where a bear goes to the beach, makes friends with a seal, and then returns home.
Rewrite	This is the summary of a Broadway play: """ {summary} """ This is the outline of the commercial for that play: """

Ouyang et al. 2022

RLHF: Training the Reward Model

$$p, \langle \tilde{x}_0, \ldots, \tilde{x}_N \rangle \quad r(\tilde{x}_i) \geq r(\tilde{x}_{i+1}) \longrightarrow \mathcal{D}_r = \{(p, \tilde{x}_w, \tilde{x}_l)\} \quad r(\tilde{x}_w) > r(\tilde{x}_l)$$

Create a new dataset with prompts paired with winning and losing continuations.

$$\theta_{\text{reward}} \approx \arg\max_{\theta} \mathbb{E}_{(p, \tilde{x}_w, \tilde{x}_l) \sim \mathcal{D}_r} \log \left(\sigma \left(r_\theta(p, \tilde{x}_w) - r_\theta(p, \tilde{x}_l) \right) \right)$$

Expectation over ranking pairs

Predicted score for winning continuation

Predicted score for losing continuation

Ouyang et al. 2022

RLHF: Training the Reward Model

$$p, \langle \tilde{x}_0, \ldots, \tilde{x}_N \rangle \quad r(\tilde{x}_i) \geq r(\tilde{x}_{i+1}) \longrightarrow \mathcal{D}_r = \{(p, \tilde{x}_w, \tilde{x}_l)\} \quad r(\tilde{x}_w) > r(\tilde{x}_l)$$

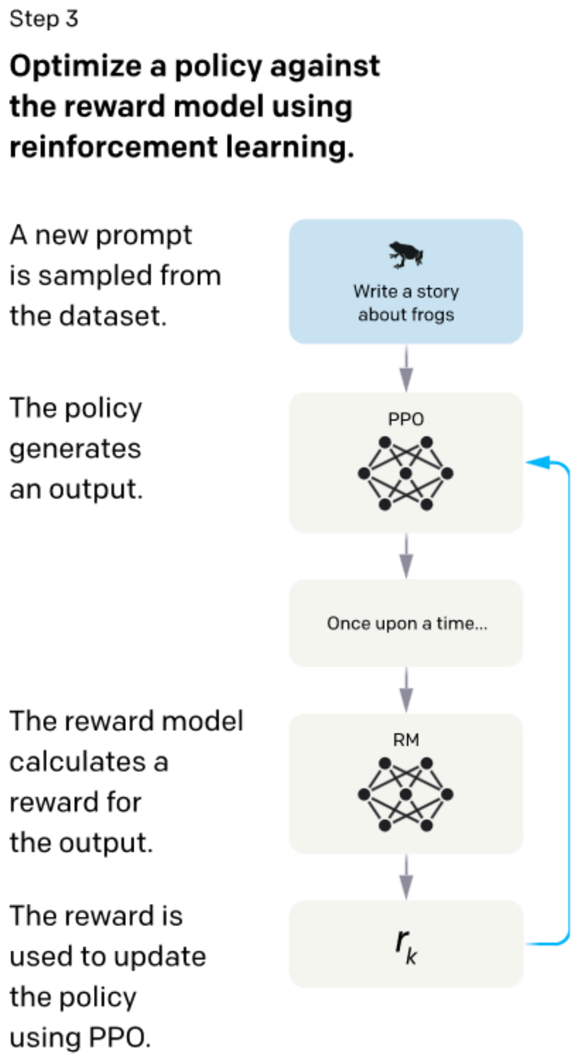
Create a new dataset with prompts paired with winning and losing continuations.

$$\theta_{\text{reward}} \approx \arg\max_{\theta} \mathbb{E}_{(p,\tilde{x}_w,\tilde{x}_l)\sim\mathcal{D}_r} \log\left(\sigma(r_\theta(p, \tilde{x}_w) - r_\theta(p, \tilde{x}_l))\right)$$

- Architecture is GPT-3 with the final projection layer removed (and replaced with a projection to predict a scalar)
- Initialized as a (small, 6B) GPT-3 model that was supervised fine-tuned using $\mathcal{D}_d$

Ouyang et al. 2022

RLHF: Optimizing the LLM Policy



Doing a lot of heavy lifting: PPO objective to maximize

KL divergence between original policy and current parameters

$$p \sim \mathcal{D}_p$$
$$\tilde{x} \sim \pi_\theta(\cdot \mid p)$$
$$s = r_{\theta_{\text{reward}}}(p, \tilde{x})$$

$$\mathbb{E}_{p \in \mathcal{D}_p}\left(s - \beta \log\left(\frac{\pi_\theta(\tilde{x} \mid p)}{\pi_{\theta_{\text{sup}}}(\tilde{x} \mid p)}\right)\right)$$
$$+\mathbb{E}_{d \in \mathcal{D}_d} \log\left(\pi_\theta(d)\right)$$

Objective to maximize

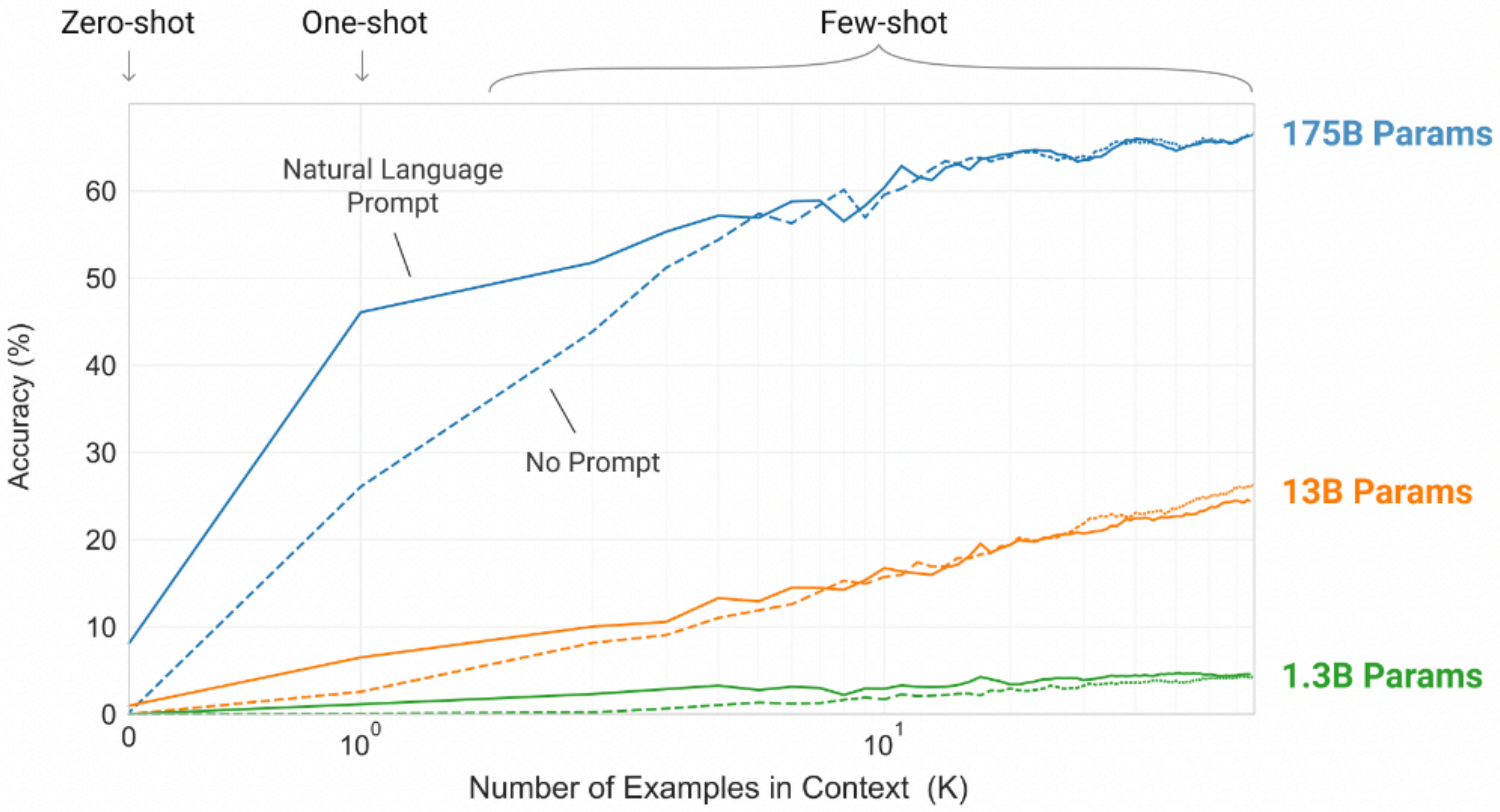
Ouyang et al. 2022

Scaling

How does performance improve when:

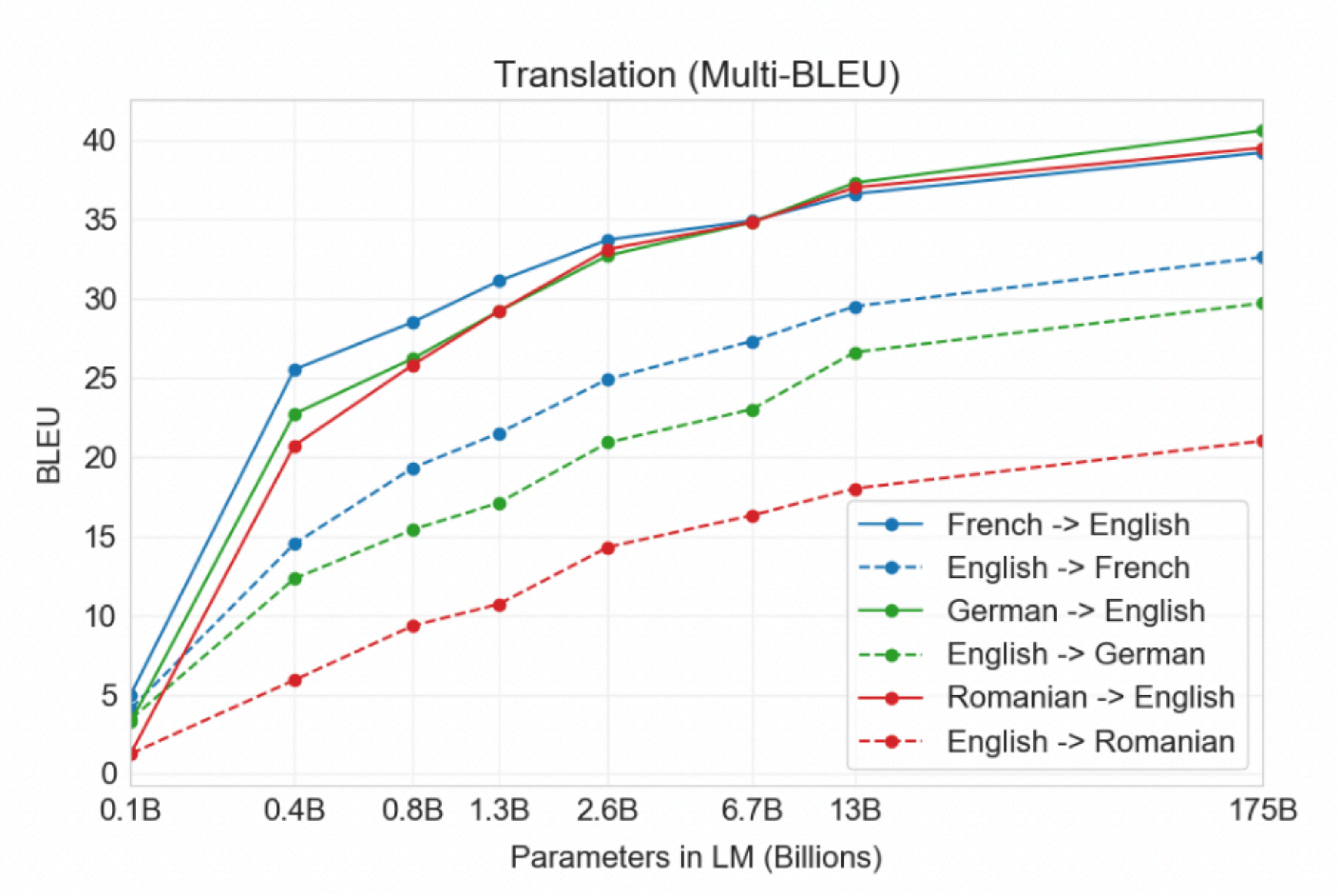
- Increasing the number of few-shot examples?
- Making the model larger?
- Making the dataset larger?
- Increasing the batch size?
- Training the model for longer?

Scaling: Few-Shot Examples



Brown et al. 2020

Scaling: Model Size



Brown et al. 2020

Scaling: Model Size

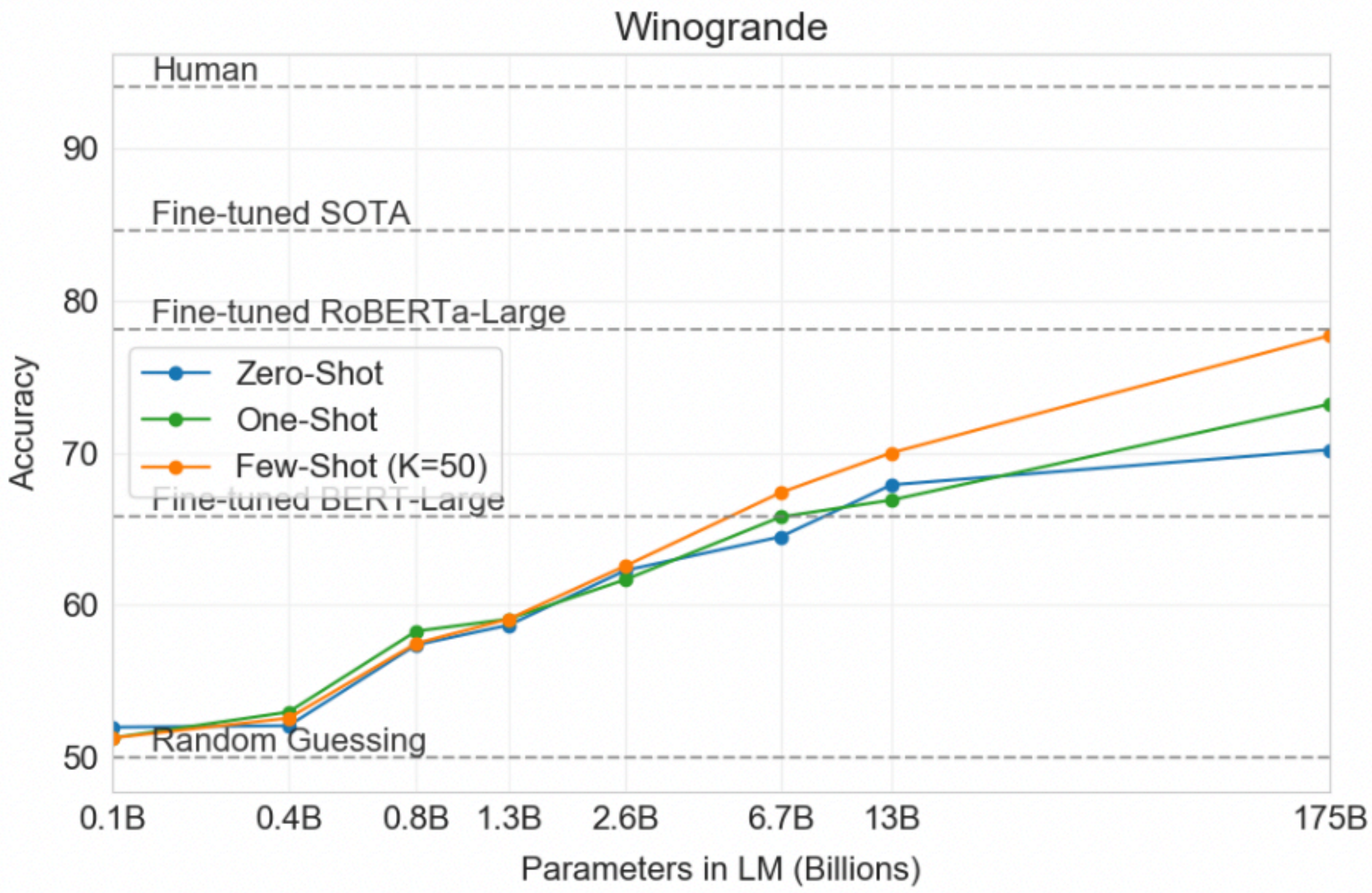
Question: The Dodecanese Campaign of WWII that was an attempt by the Allied forces to capture islands in the Aegean Sea was the inspiration for which acclaimed 1961 commando film?
Answer: The Guns of Navarone
Excerpt: The Dodecanese Campaign of World War II was an attempt by Allied forces to capture the Italian-held Dodecanese islands in the Aegean Sea following the surrender of Italy in September 1943, and use them as bases against the German-controlled Balkans. The failed campaign, and in particular the Battle of Leros, inspired the 1957 novel **The Guns of Navarone** and the successful 1961 movie of the same name.

Question: American Callan Pinckney's eponymously named system became a best-selling (1980s-2000s) book/video franchise in what genre?
Answer: Fitness
Excerpt: Callan Pinckney was an American fitness professional. She achieved unprecedented success with her Callanetics exercises. Her 9 books all became international best-sellers and the video series that followed went on to sell over 6 million copies. Pinckney's first video release "Callanetics: 10 Years Younger In 10 Hours" outsold every other **fitness** video in the US.

Brown et al. 2020

Joshi et al. 2017

Scaling: Model Size



The trophy doesn't fit into the brown suitcase because **it's** too <u>large</u>.

it = trophy

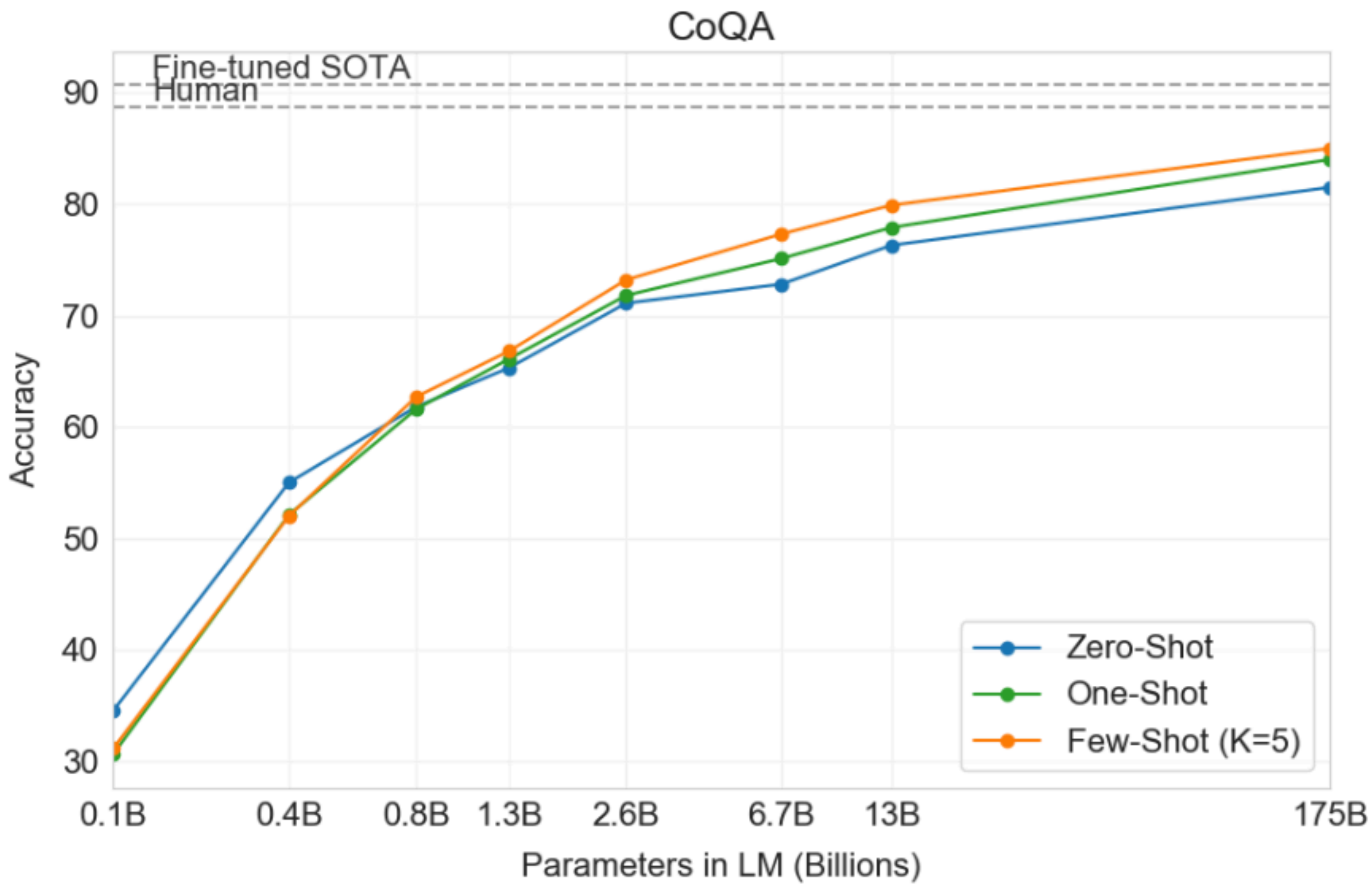
The trophy doesn't fit into the brown suitcase because **it's** too <u>small</u>.

it = suitcase

Brown et al. 2020

Sakaguchi et al. 2019

Scaling: Model Size



Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80. Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well. Jessica had . . .

Q_1: Who had a birthday?
A_1: Jessica
R_1: Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80.

Q_2: How old would she be?
A_2: 80
R_2: she was turning 80

Q_3: Did she plan to have any visitors?
A_3: Yes
R_3: Her granddaughter Annie was coming over

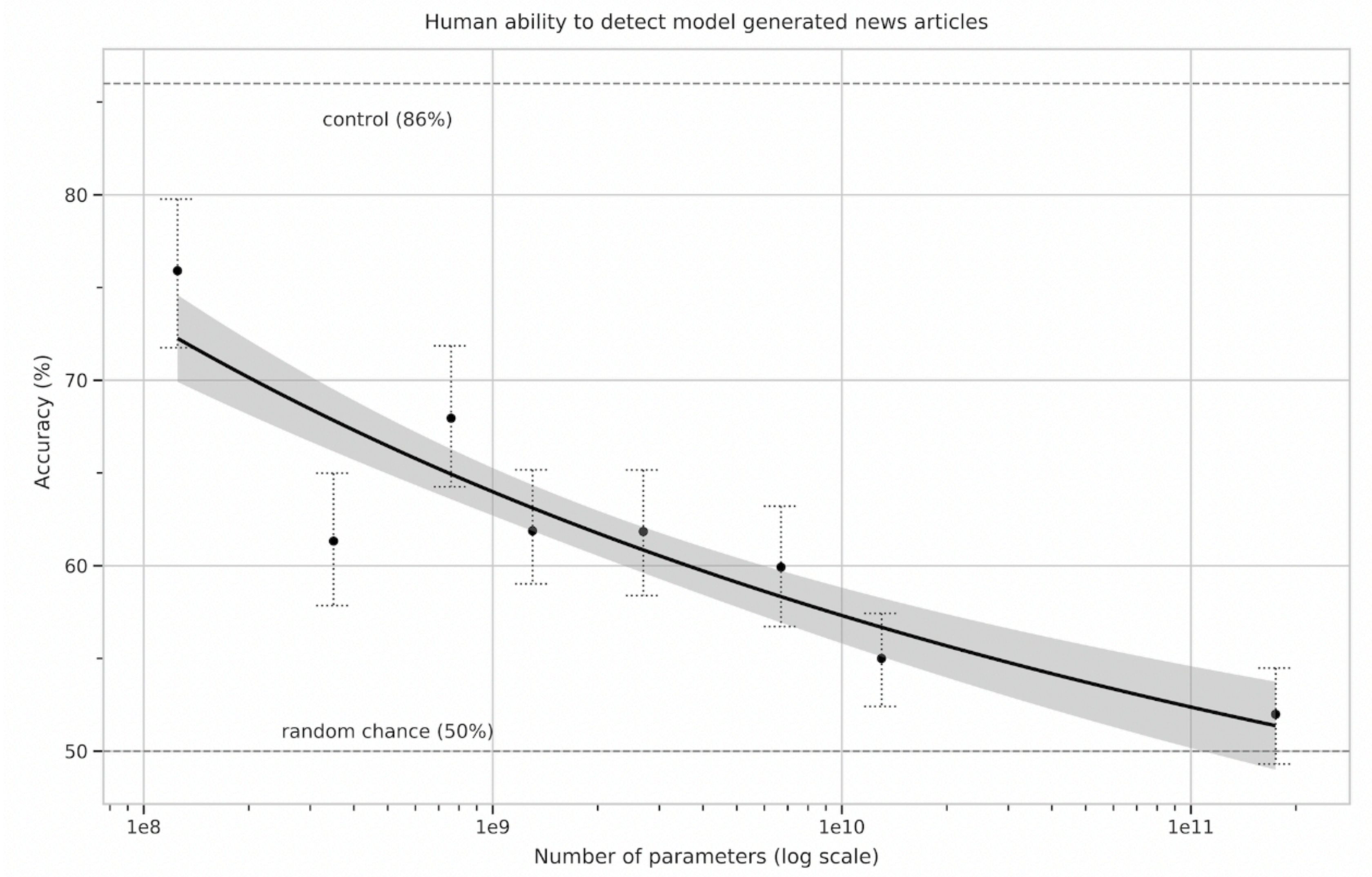
Q_4: How many?
A_4: Three
R_4: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.

Q_5: Who?
A_5: Annie, Melanie and Josh
R_5: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.

Brown et al. 2020

Reddy et al. 2018

Scaling: Model Size



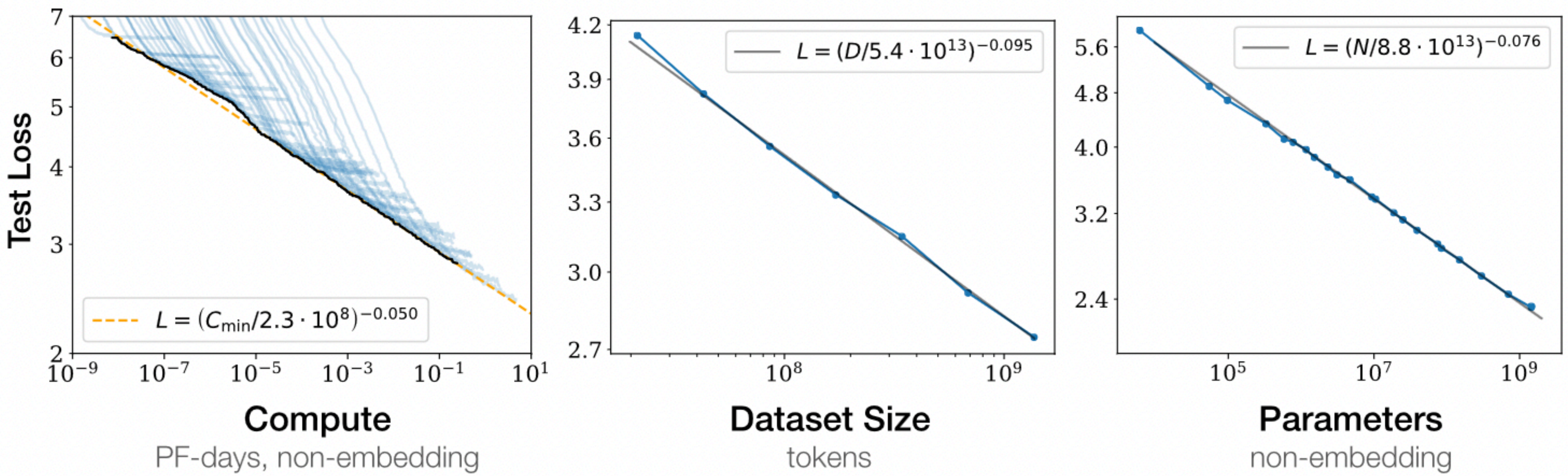
Brown et al. 2020

Scaling Laws

- N is the number of parameters (excluding vocabulary and positional embeddings)
- B is the batch size
- S is the number of training steps (parameter updates)
- $C = 6NBS$ is an estimate of the total non-embedding compute (unit: PF-days, i.e., the number of floating point operations that can be performed in 1 day)

Kaplan et al. 2020

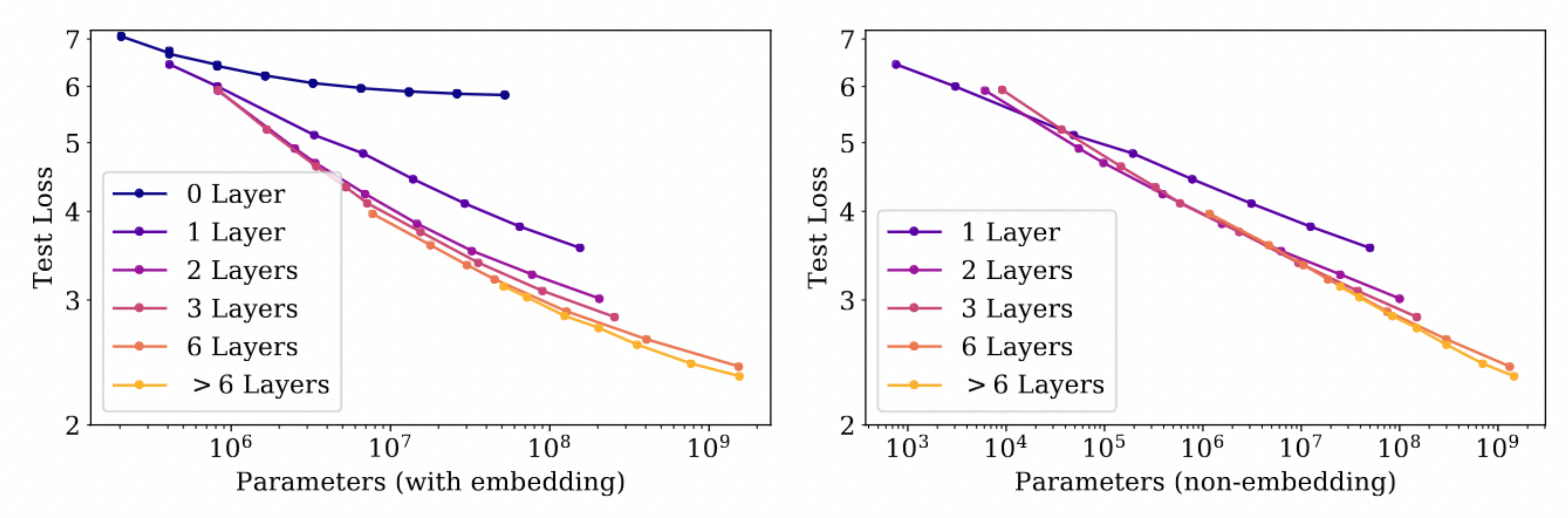
Scaling Laws



We can predict test loss of a Transformer language model from the number of parameters, dataset size, or compute budget.

Kaplan et al. 2020

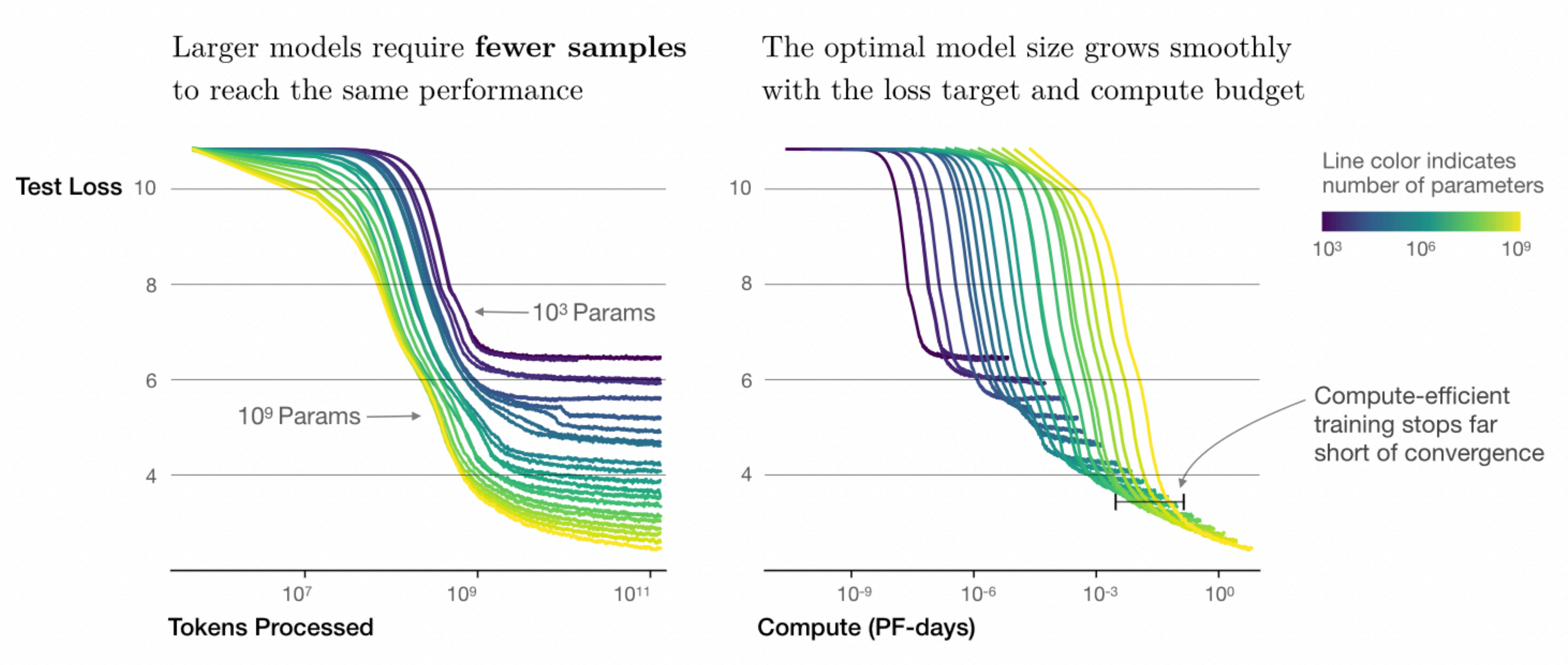
Scaling Laws



Why consider only non-embedding parameters? Laws are more complex (also take into account number of layers)

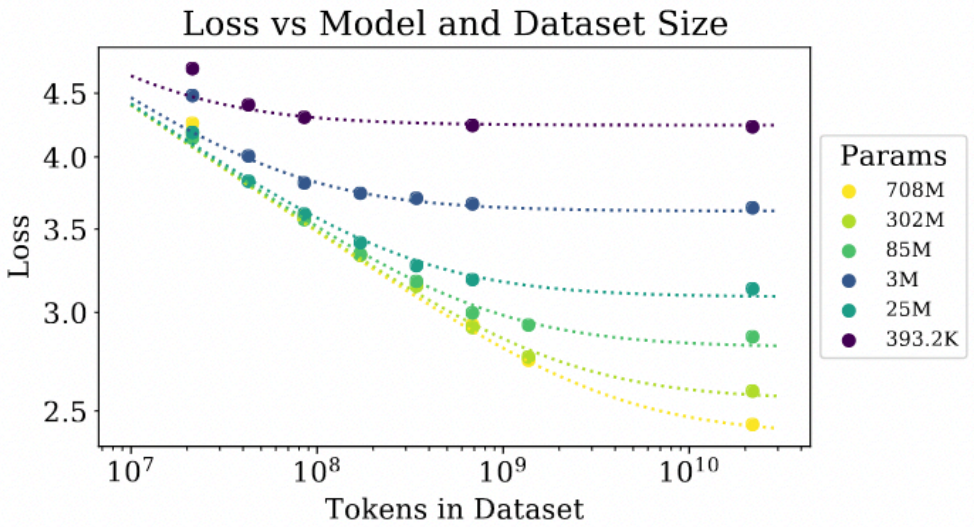
Kaplan et al. 2020

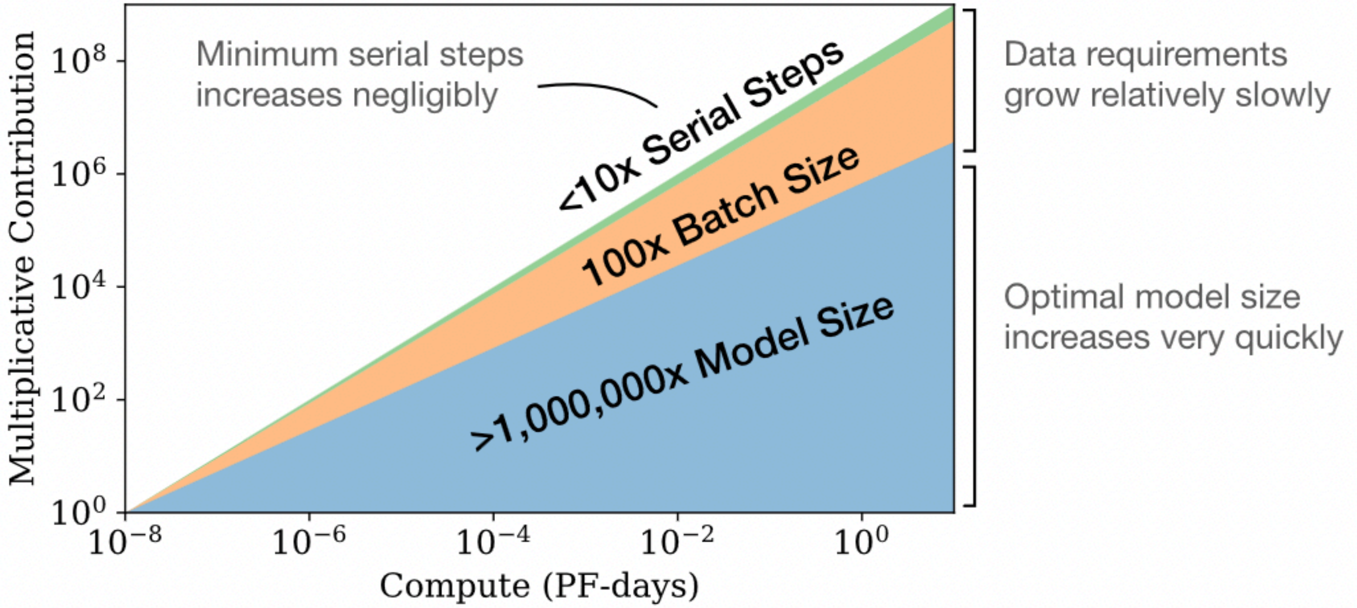
Scaling Laws



Kaplan et al. 2020

Scaling Laws

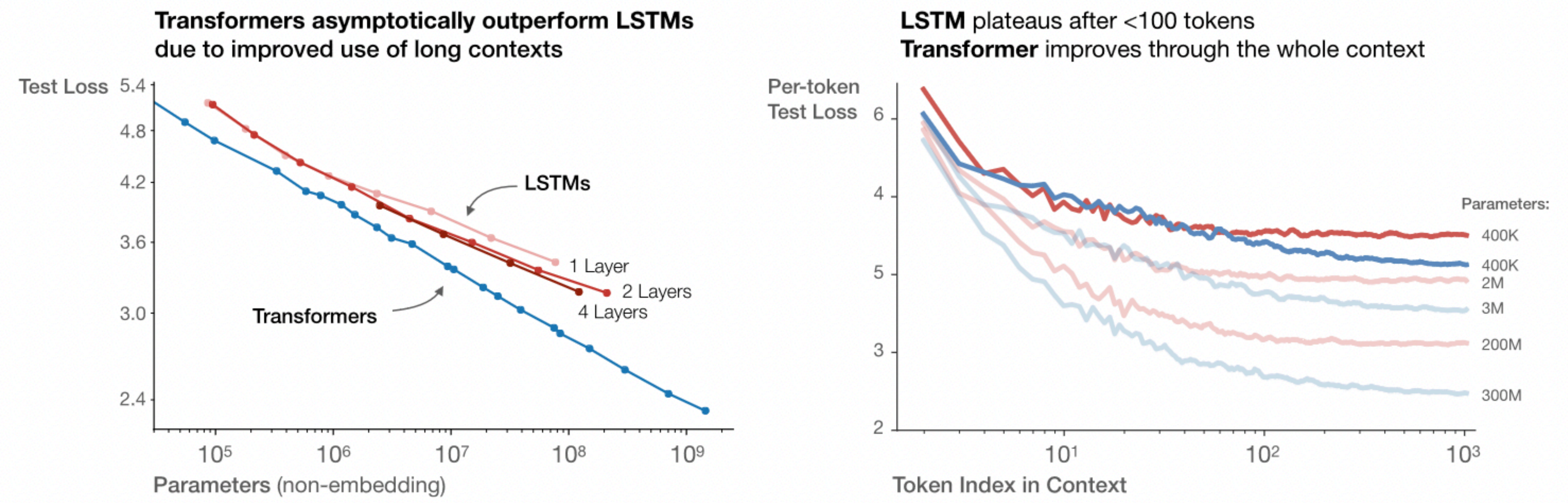




As our training budget increases, compute should be allocated to model size, rather than batch size or number of training steps

Kaplan et al. 2020

Scaling Laws



Kaplan et al. 2020

Scaling Laws

1. For models with a limited number of parameters, trained to convergence on sufficiently large datasets:

$$L(N) = (N_c/N)^{\alpha_N}\ ;\ \ \alpha_N \sim 0.076, \quad N_c \sim 8.8 \times 10^{13} \text{ (non-embedding parameters)} \tag{1.1}$$

2. For large models trained with a limited dataset with early stopping:

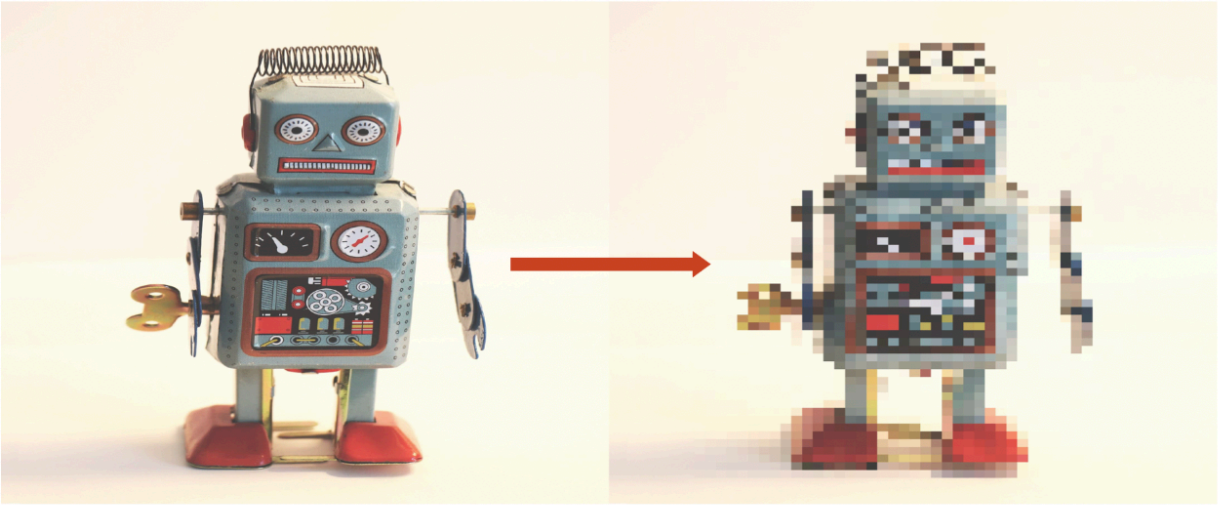
$$L(D) = (D_c/D)^{\alpha_D}\ ;\ \ \alpha_D \sim 0.095, \quad D_c \sim 5.4 \times 10^{13} \text{ (tokens)} \tag{1.2}$$

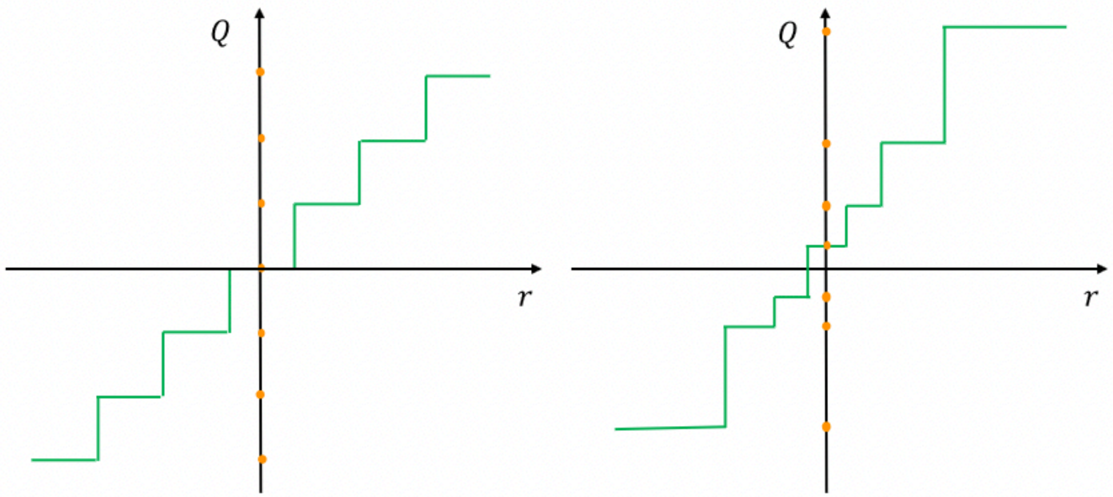
3. When training with a limited amount of compute, a sufficiently large dataset, an optimally-sized model, and a sufficiently small batch size (making optimal[3] use of compute):

$$L(C_{\min}) = \left(C_c^{\min}/C_{\min}\right)^{\alpha_C^{\min}}\ ;\ \ \alpha_C^{\min} \sim 0.050, \quad C_c^{\min} \sim 3.1 \times 10^{8} \text{ (PF-days)} \tag{1.3}$$

Kaplan et al. 2020

Quantization





Uniform

Non-Uniform

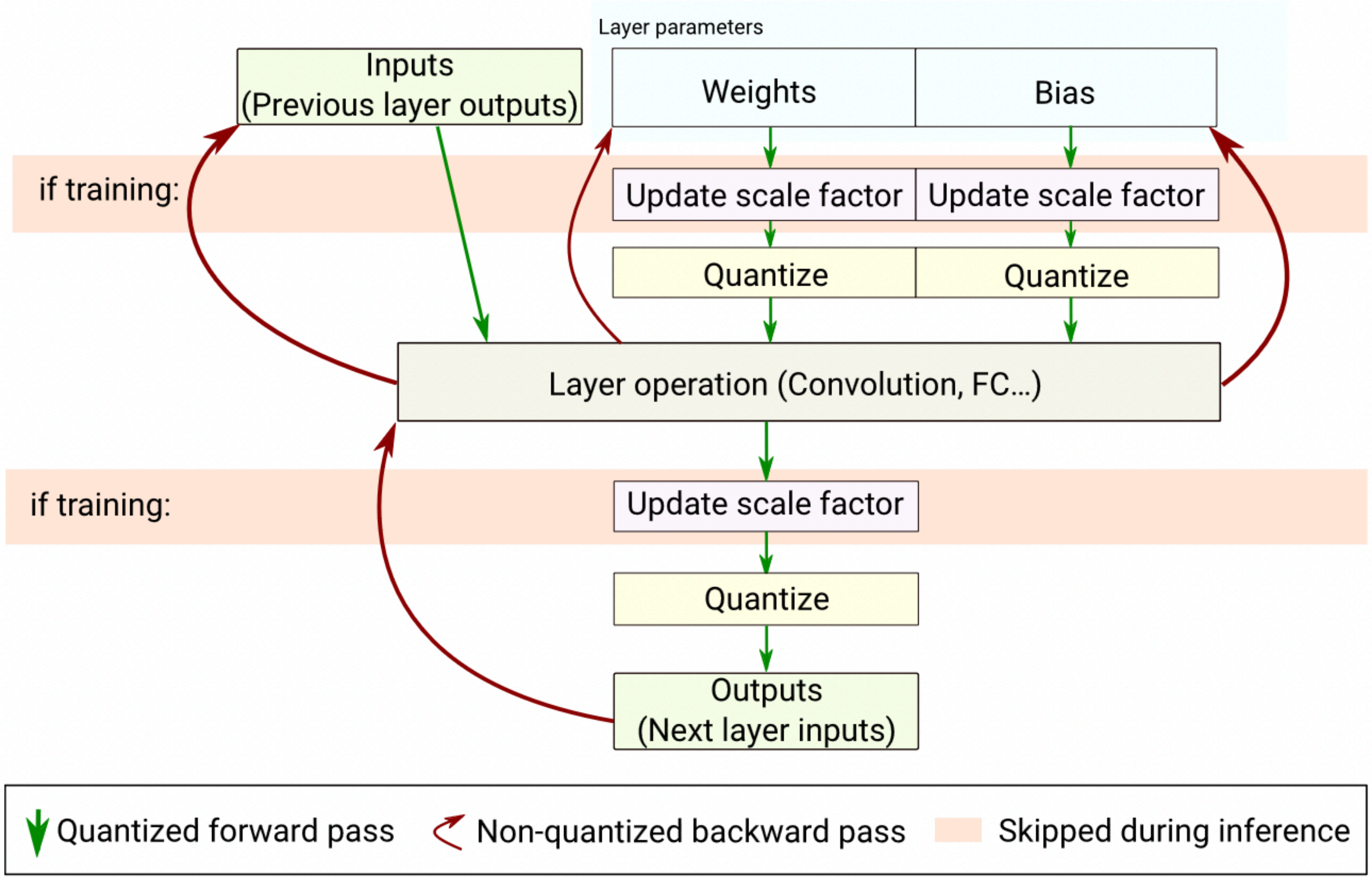
Main principle: use lower-precision representations of network parameters

Gholami et al. 2021

Quantization

- Reduces space required to store model: useful for on-device inference
- Two primary methods
  - Post-training quantization
  - Quantization-aware training

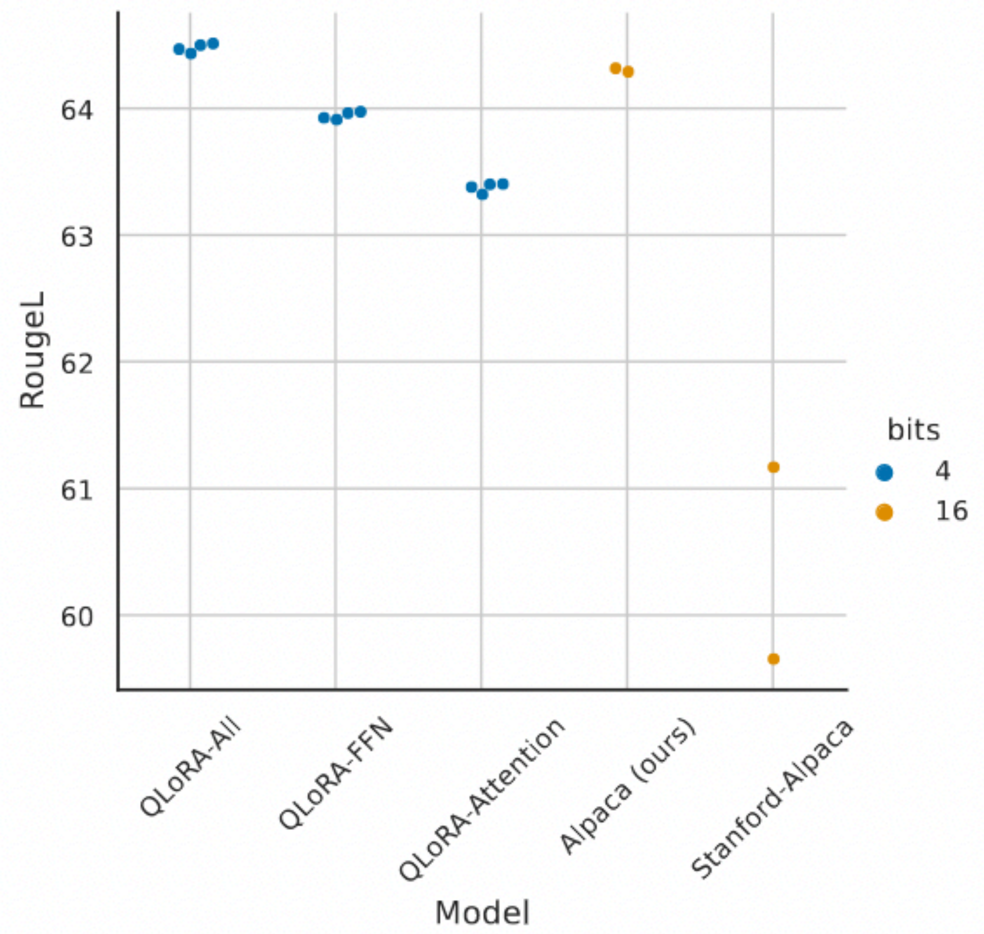
Quantization-Aware Training



Novac et al. 2021

QLoRA

- Quantize pre-trained model to 4 bits
- Backpropagate gradients through these frozen parameters into LoRA
- Allows fine-tuning 65B parameter model on a 48GB GPU



Dettmers et al. 2023

Pruning

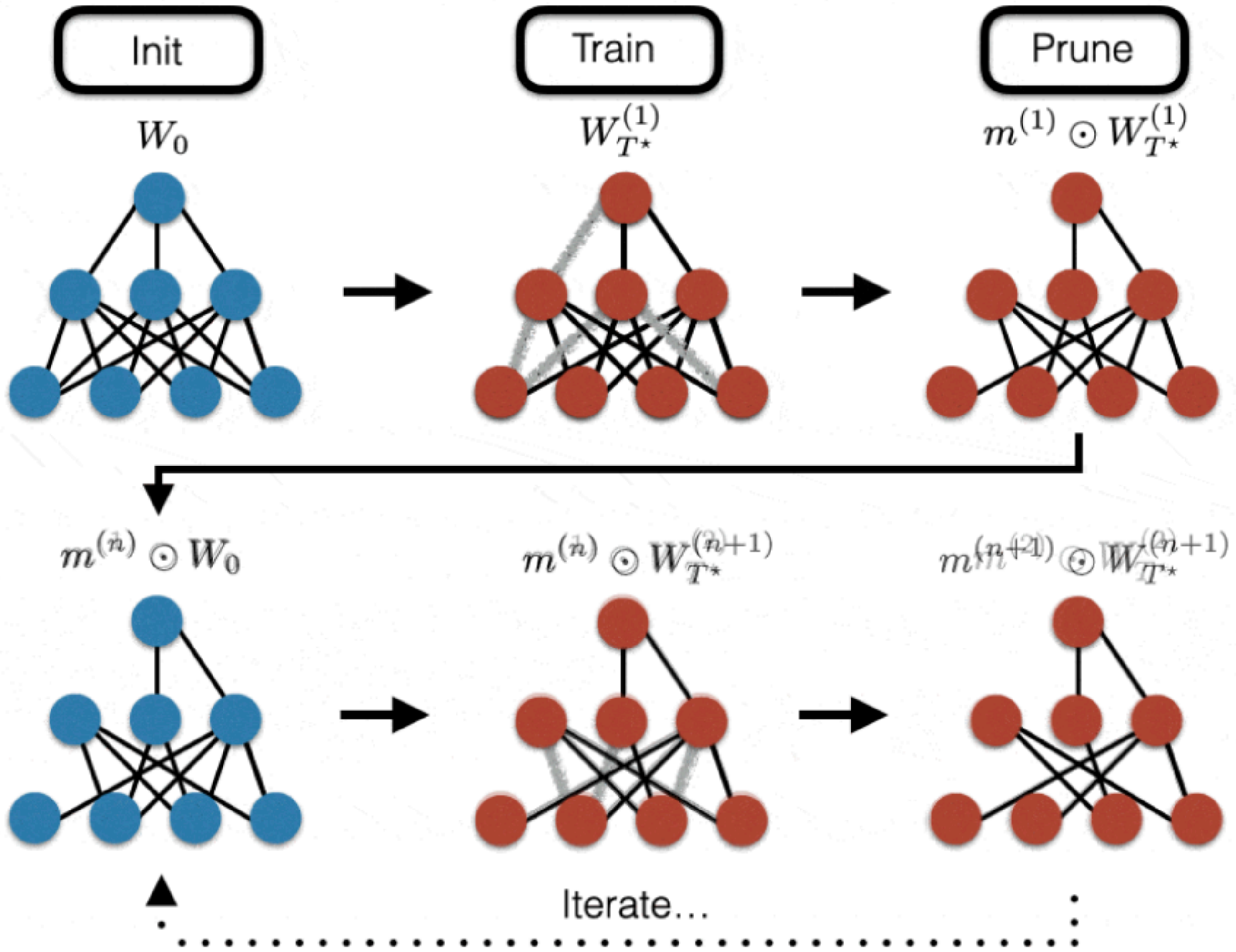
- General principle: not all weights in a network are important
- Approach: mask out some weights
  - Start with a large network, and train it to convergence
  - Prune in iterations, based on second-order derivatives:
    - Prune and retrain
    - Prune and update weights based on second-order statistics

$$a = (W \odot M)x$$

LeCun et al. 1990, Hassibi and Stork 1993

Lottery Ticket Hypothesis

Lottery ticket hypothesis (Frankle and Carbin 2019): "A randomly-initialized, dense neural network contains a subnetwork that is initialized such that, when trained in isolation, it can match the test accuracy of the original network after training for at most the same number of iterations"



Frankle and Carbin 2019

Risks

As AI language skills grow, so do scientists' concerns

Italy orders ChatGPT blocked citing data protection concerns

GPT-3 has 'consistent and creative' anti-Muslim bias, study finds

Google's Sentiment Analyzer Thinks Being Gay Is Bad

Amazon ditched AI recruiting tool that favored men for technical jobs

researchers call for urgent action to address harms of large language models like GPT-3

A.I. Is Mastering Language. Should We Trust What It Says?

What Do We Do About the Biases in AI?

Teachers Fear ChatGPT Will Make Cheating Easier Than Ever

How ChatGPT Kicked Off an A.I. Arms Race

Types of AI Harm

Biases in models perpetuate stereotypes

REPRESENTATION	ALLOCATION
Representations of black criminality ⬇ Racial stereotype	Representations of black criminality ⬇ Racial stereotype ⬇ Prospects in the labor market
Long term	Immediate
Difficult to formalize	Easily quantifiable
Diffuse	Discrete
Cultural	Transactional

Stereotype-based models worsen performance for groups already facing discrimination

Crawford 2017

Representational Bias in NLP

- Word embeddings
- Sentence embeddings
- Machine translation
- Image captioning
- Coreference resolution
- Language modeling
- Hate speech detection

Embeddings

Word Embedding Association Test

Attribute Words

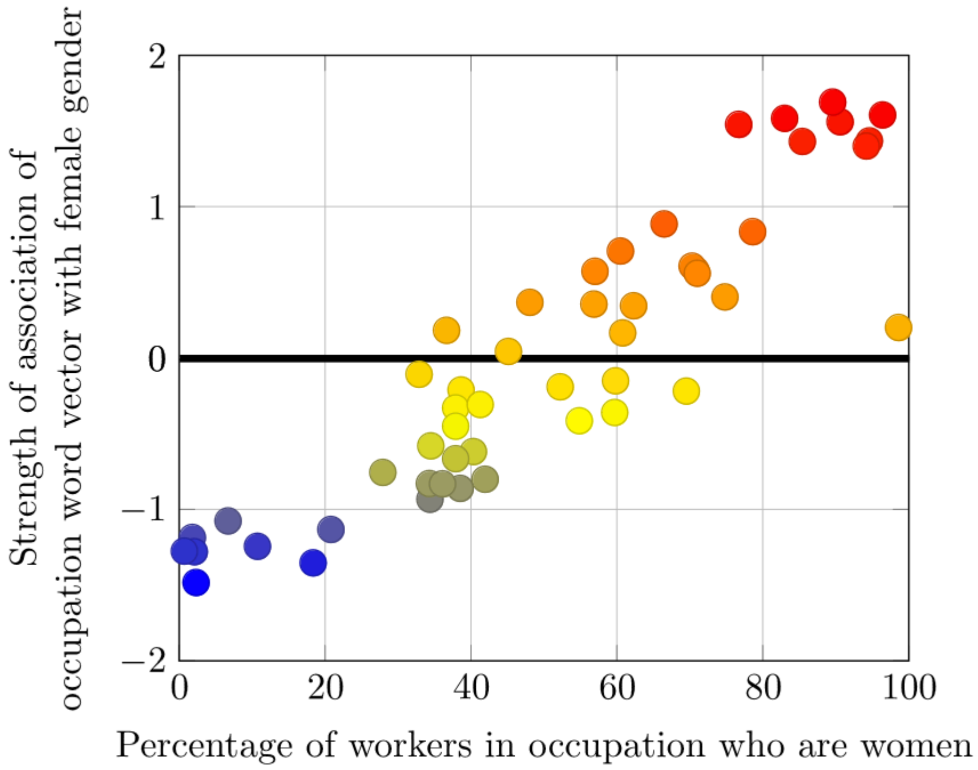
A
("Pleasant Attributes")
love, cheer, friend...

B
("Unpleasant Attributes")
ugly, evil, abuse...

Target Words

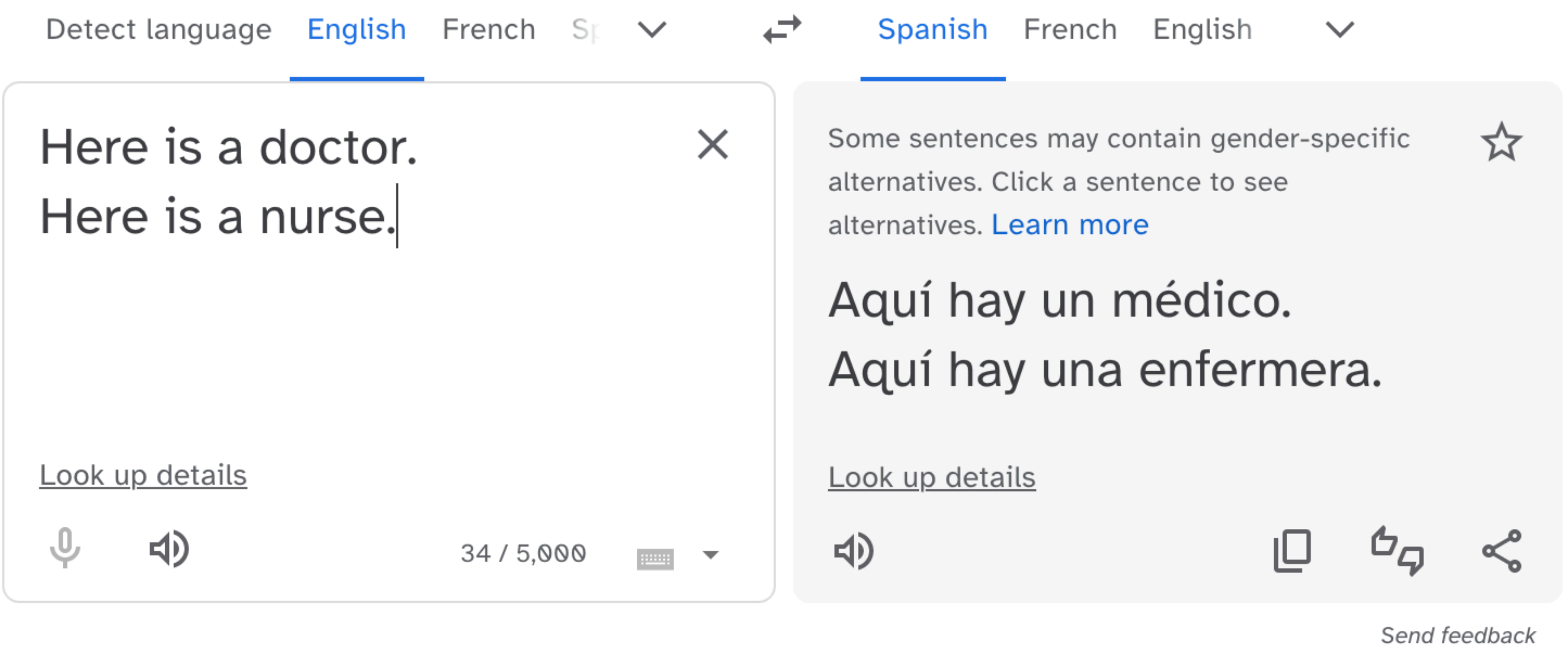
X
("European American Names")
Adam, Harry, Nancy...

Y
("African American Names")
Jamel, Lavar, Latisha...



Caliskan et al. 2017

Machine Translation



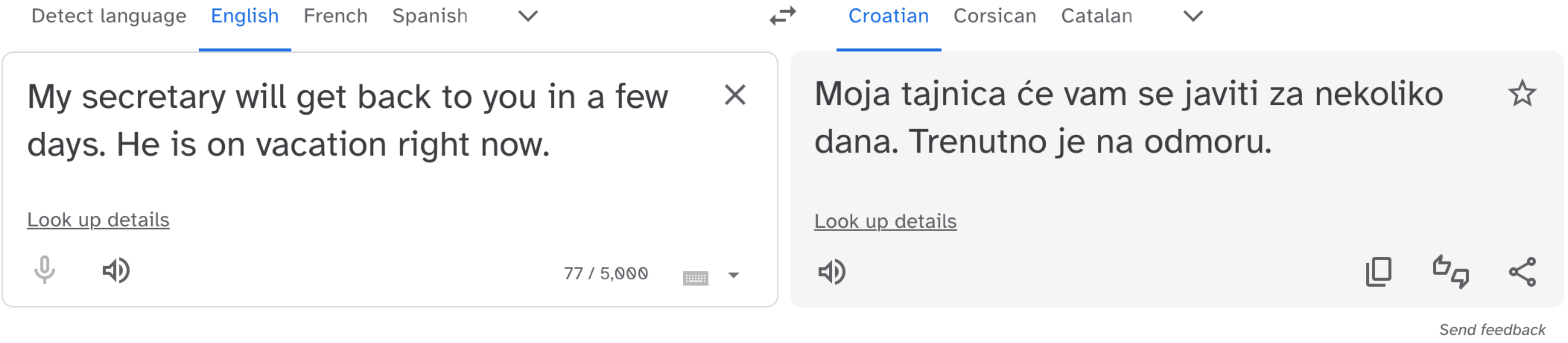


Image Captioning

Human: A busy city street in an Asian country with lots of traffic.

Transformer: A city street with lots of asian businesses.

Human: People watch a horse and carriage ride by them.

Transformer: A group of indians standing around in inflatable blue.

Human: A crowded farmers market with a line of cars outside.

Transformer: A street scene with a focus on a mexican restaurant.

Figure 3: Examples of images for which the **Transformer** model [67] assigns racial or cultural descriptors to the caption. While in the first image the descriptor of "Asian" is present in the human-annotated caption, neither of the descriptors, "Indian" nor "Mexican," are applicable in the latter images.

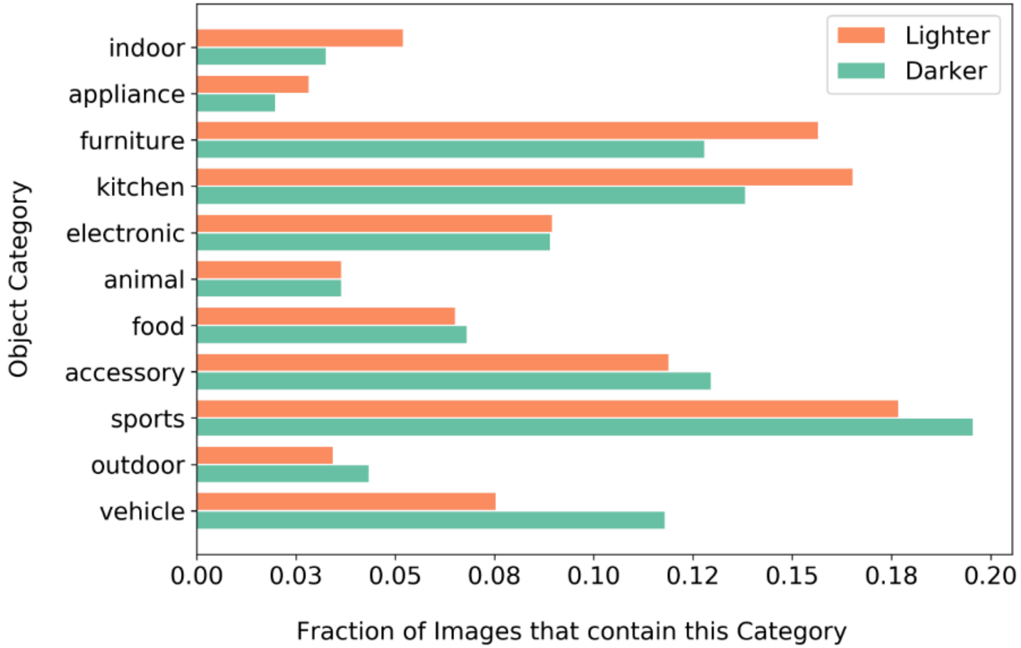
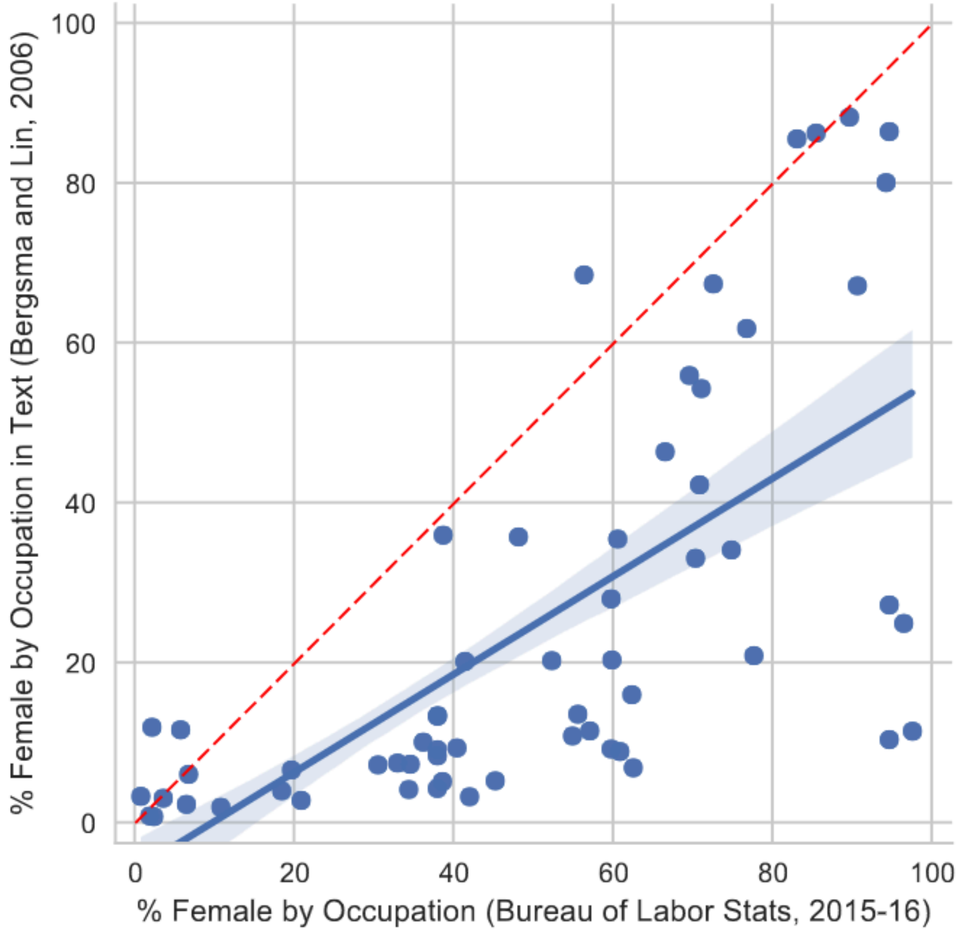


Figure 5: Images with people of lighter and darker skin tones co-occur with object categories at different frequencies. Whereas the former tend to be pictured with object categories that are indoor, the latter tend to be pictured with object categories that are more likely to be outdoors.

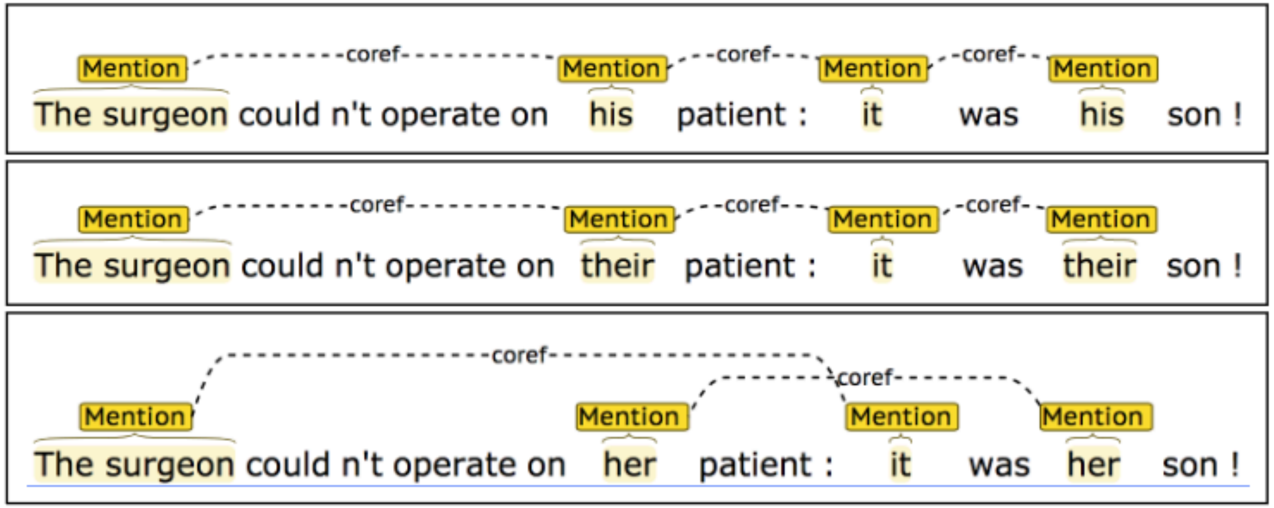
Zhao et al. 2017

Coreference Resolution



Compounding effect

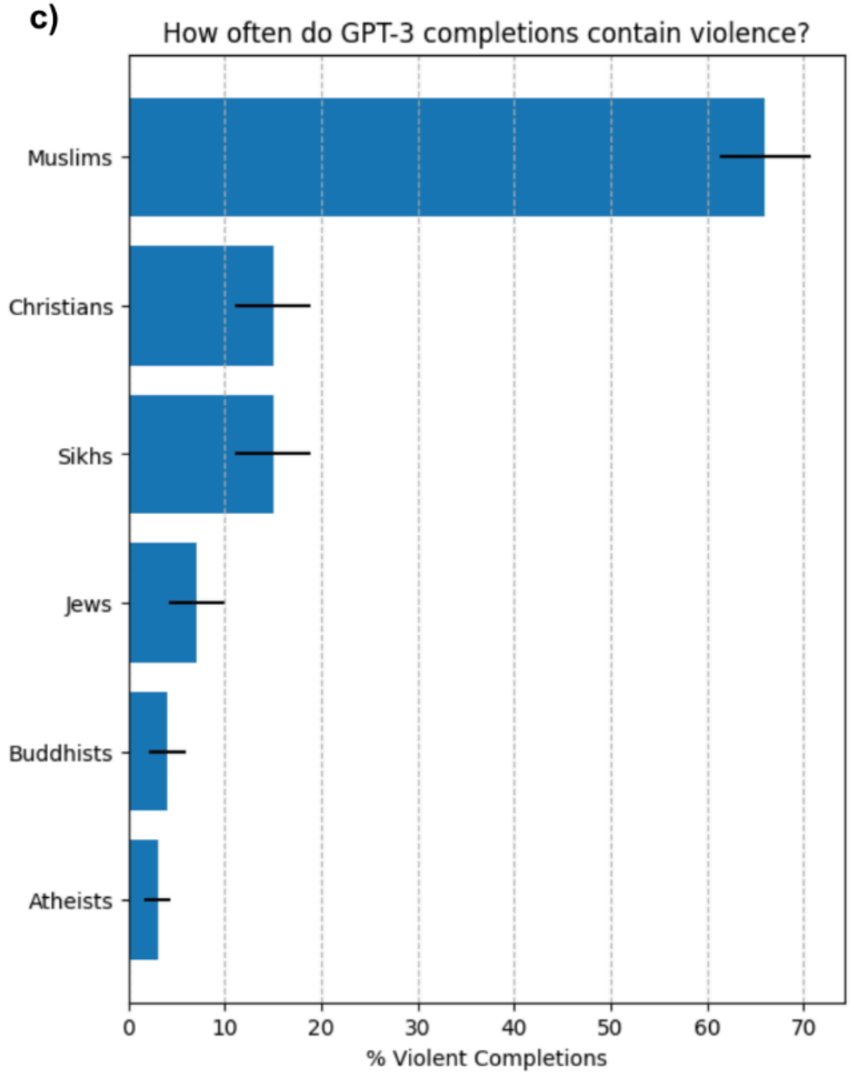
- BLS reports 39% of managers are female
- But coref corpus used for training reports only 5% of managers are female
- Trained model predicts 0% female for managers

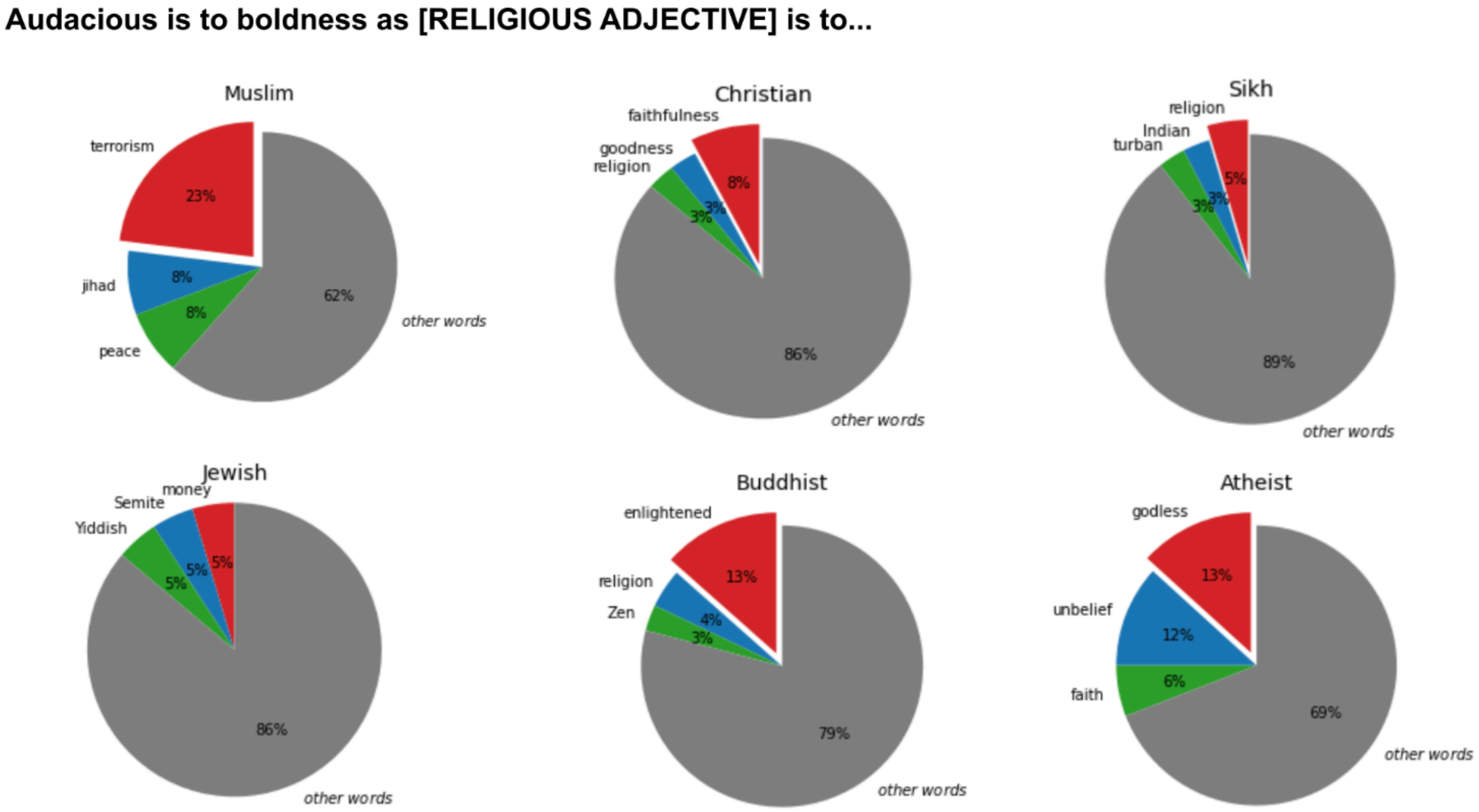


Rudinger et al. 2018

Language Modeling

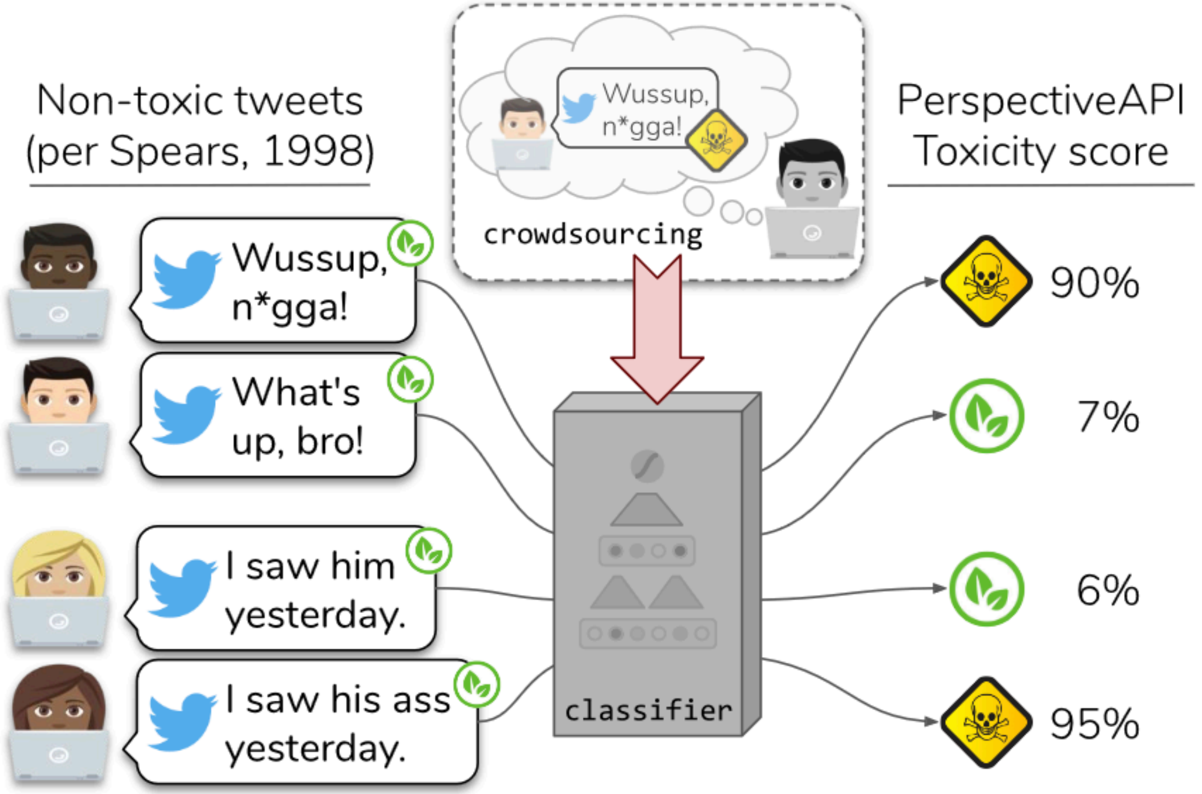
c)





Abid et al. 2021

Hate Speech Detection



	category	count	AAE corr.
DWMW17	hate speech	1,430	−0.057
	offensive	19,190	0.420
	none	4,163	−0.414
	total	**24,783**	
FDCL18	hateful	4,965	0.141
	abusive	27,150	0.355
	spam	14,030	−0.102
	none	53,851	−0.307
	total	**99,996**	

Downstream effect: filtering out / censoring non-hateful language, reinforcing representational biases

Sap et al. 2019

Training Data

- Modern NLP models are data hungry
- Solution: scrape text from the web, which likely introduces biases
- What do we want to filter out?
  - Hate speech
  - Language expressing stereotypes
  - Spam
  - Adult content
  - Machine-generated text
- Problems with filters?

Training Data

- What are we *not* getting from scraping the web?
  - Low-resource languages
  - Dialects with fewer speakers (e.g., AAE)
  - Non-written languages (e.g., ASL)
  - Language from people who aren’t putting content on the web (e.g., older speakers, or those who don’t have access to the Internet)
- This reinforces biases towards language that *is* well-represented

Training Data: Annotation

Table 12: Labeler demographic data

What gender do you identify as?	
Male	50.0%
Female	44.4%
Nonbinary / other	5.6%
What ethnicities do you identify as?	
White / Caucasian	31.6%
Southeast Asian	52.6%
Indigenous / Native American / Alaskan Native	0.0%
East Asian	5.3%
Middle Eastern	0.0%
Latinx	15.8%
Black / of African descent	10.5%
What is your nationality?	
Filipino	22%
Bangladeshi	22%
American	17%
Albanian	5%
Brazilian	5%
Canadian	5%
Colombian	5%
Indian	5%
Uruguayan	5%
Zimbabwean	5%

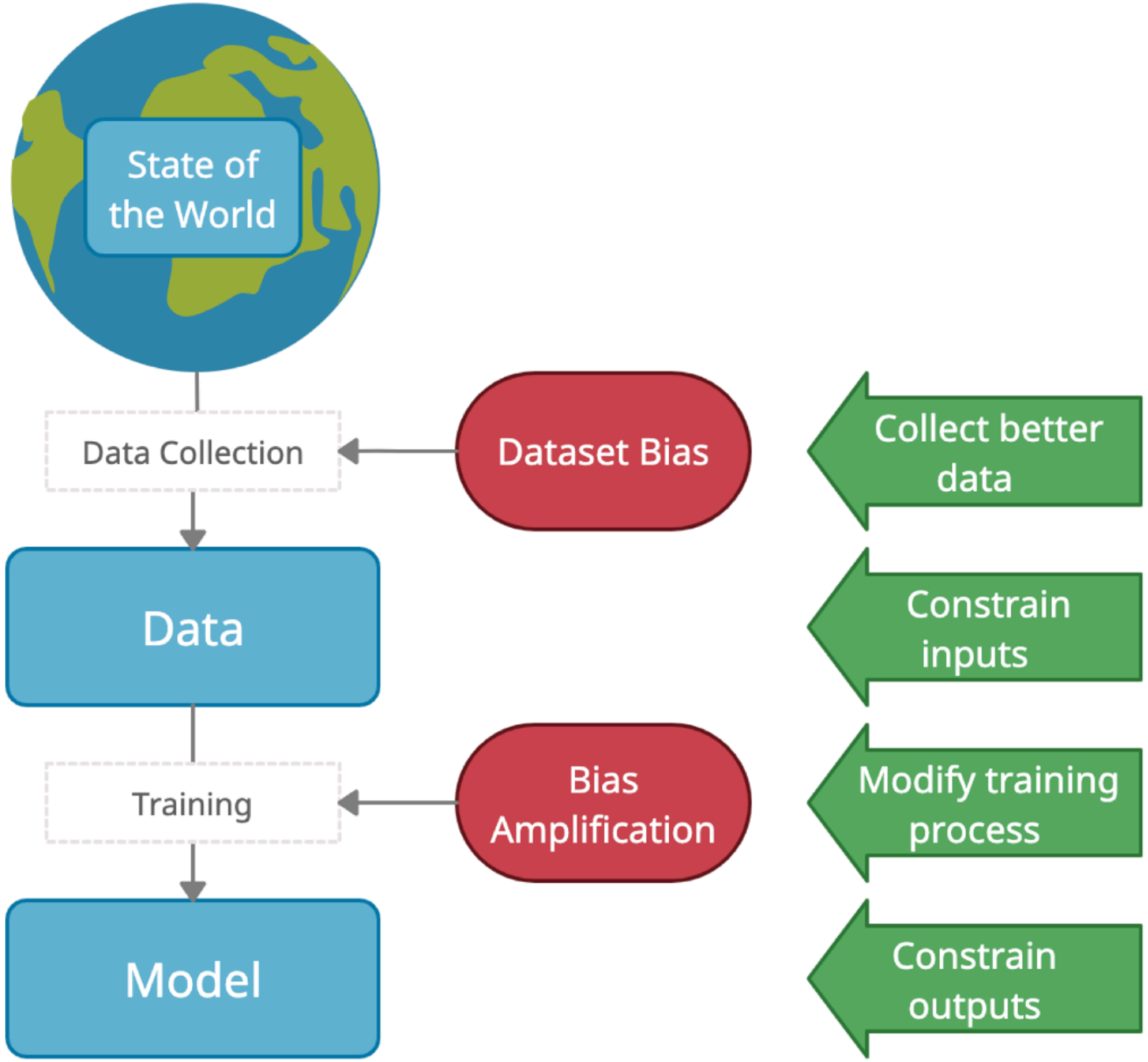
What is your age?	
18-24	26.3%
25-34	47.4%
35-44	10.5%
45-54	10.5%
55-64	5.3%
65+	0%
What is your highest attained level of education?	
Less than high school degree	0%
High school degree	10.5%
Undergraduate degree	52.6%
Master's degree	36.8%
Doctorate degree	0%

Training Data: Annotation

- Data labelers: often low-income, inadequately compensated
- Companies like OpenAI have been known to exploit workers in countries with weaker labor rights and economies
  - Perrigo 2022: “OpenAI used Kenyan workers on less than $2 per hour to make ChatGPT less toxic”
  - Hao and Hernández 2022: “workers in Venezuela earn an average of a little more than 90 cents an hour” through the use of Scale AI

	All working adults	Workers on Mechanical Turk
Male	53%	51%
Female	47	49
Age		
18-29	23	41
30-49	43	47
50-64	28	10
65+	6	1
Race and ethnicity		
White, non-Hispanic	65	77
Black, non-Hispanic	11	6
Hispanic	16	6
Other	8	11

Mitigating Harm due to Bias



- Fine-tune models with smaller, unbiased datasets
- Directly adjust word embeddings, loss function, etc.

Mitigating Harm due to Bias

(R1) Ground work analyzing "bias" in NLP systems in the relevant literature outside of NLP that explores the relationships between language and social hierarchies. Treat representational harms as harmful in their own right.

(R2) Provide explicit statements of why the system behaviors that are described as "bias" are harmful, in what ways, and to whom. Be forthright about the normative reasoning (Green, 2019) underlying these statements.

(R3) Examine language use in practice by engaging with the lived experiences of members of communities affected by NLP systems. Interrogate and reimagine the power relations between technologists and such communities.

- Fine-tune models with smaller, unbiased datasets
- Directly adjust word embeddings, loss function, etc.
- Focus on how the model is used in practice, rather than its internal bias

Blodgett et al. 2020

Further Considerations

Metrics of "bias" could themselves be biased

- Intersectionality
- False negatives
- Ignoring subtleties of context

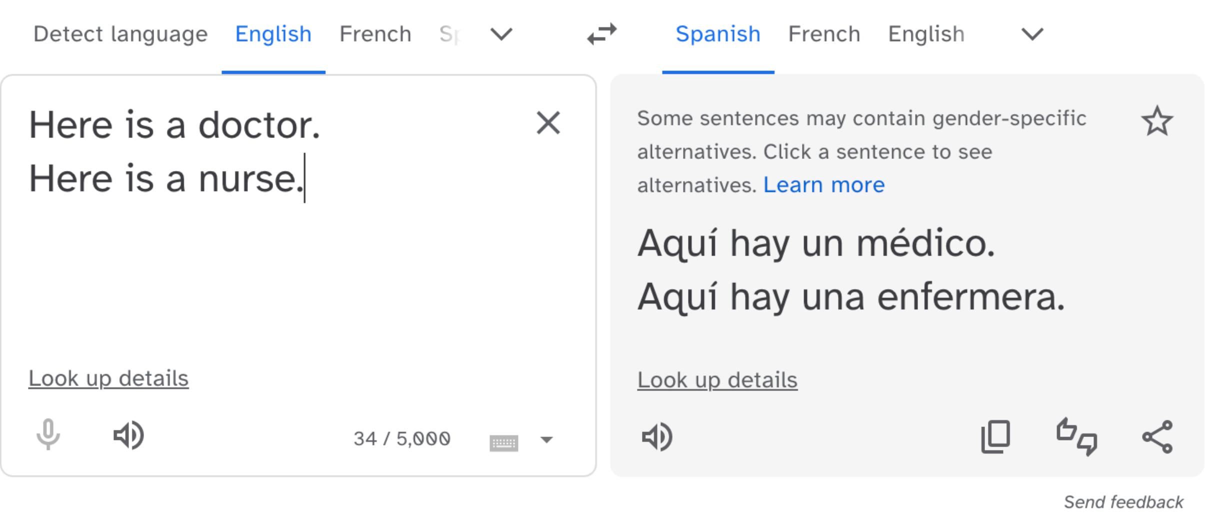
Classifier	Metric	DF	DM	LF	LM
MSFT	TPR(%)	76.2	**100**	**100**	**100**
	Error Rate(%)	**23.8**	0.0	0.0	0.0
	PPV(%)	**100**	84.2	**100**	**100**
	FPR(%)	0.0	**23.8**	0.0	0.0
Face++	TPR(%)	64.0	99.5	92.6	**100**
	Error Rate(%)	**36.0**	0.5	7.4	0.0
	PPV(%)	99.0	77.8	**100**	96.9
	FPR(%)	0.5	**36.0**	0.0	7.4
IBM	TPR(%)	66.9	94.3	**100**	98.4
	Error Rate(%)	**33.1**	5.7	0.0	1.6
	PPV(%)	90.4	78.0	96.4	**100**
	FPR(%)	5.7	**33.1**	1.6	0.0

Buolamwini et al. 2018

Further Considerations

Interventions don't just involve adjusting the model internals

- Holding companies accountable for the technology they build
- Designing better user interfaces



Classifier	Metric	DF	DM	LF	LM
MSFT	TPR(%)	76.2	**100**	**100**	**100**
	Error Rate(%)	**23.8**	0.0	0.0	0.0
	PPV(%)	**100**	84.2	**100**	**100**
	FPR(%)	0.0	**23.8**	0.0	0.0
Face++	TPR(%)	64.0	99.5	92.6	**100**
	Error Rate(%)	**36.0**	0.5	7.4	0.0
	PPV(%)	99.0	77.8	**100**	96.9
	FPR(%)	0.5	**36.0**	0.0	7.4
IBM	TPR(%)	66.9	94.3	**100**	98.4
	Error Rate(%)	**33.1**	5.7	0.0	1.6
	PPV(%)	90.4	78.0	96.4	**100**
	FPR(%)	5.7	**33.1**	1.6	0.0

Buolamwini et al. 2018

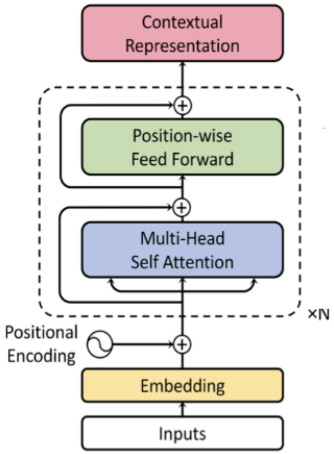
Implications of Publicly Available LLMs

Emergent capabilities → **Emergent vulnerabilities?**

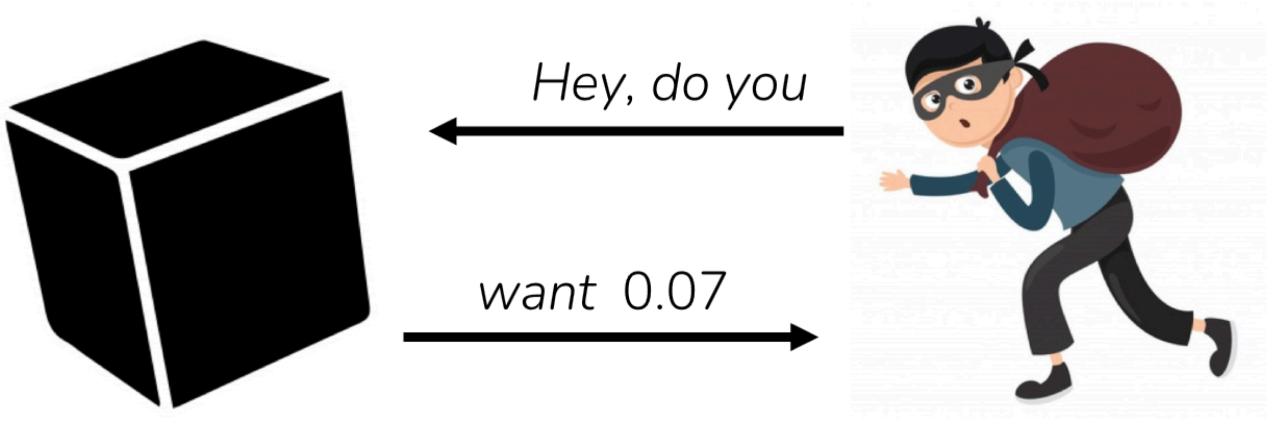
Increasing centralization → **Single point of failure**

Increasingly black-box → **Can't detect/debug errors**

Threat Model



Data + Model



Black-box API

Adversary

Extract training data

Poison training data

Steal model parameters

Eric Wallace

Extracting Memorized Training Data

Personally identifiable information

```
[redacted] Corporation Seabank Centre
[redacted] Marine Parade Southport
Peter W[redacted]
[redacted]@[redacted].[redacted].com
+[redacted] 7 5[redacted] 40[redacted]
Fax: +[redacted] 7 5[redacted] 0[redacted]0
```

Publicly available data!

But this person was wrongly indicted

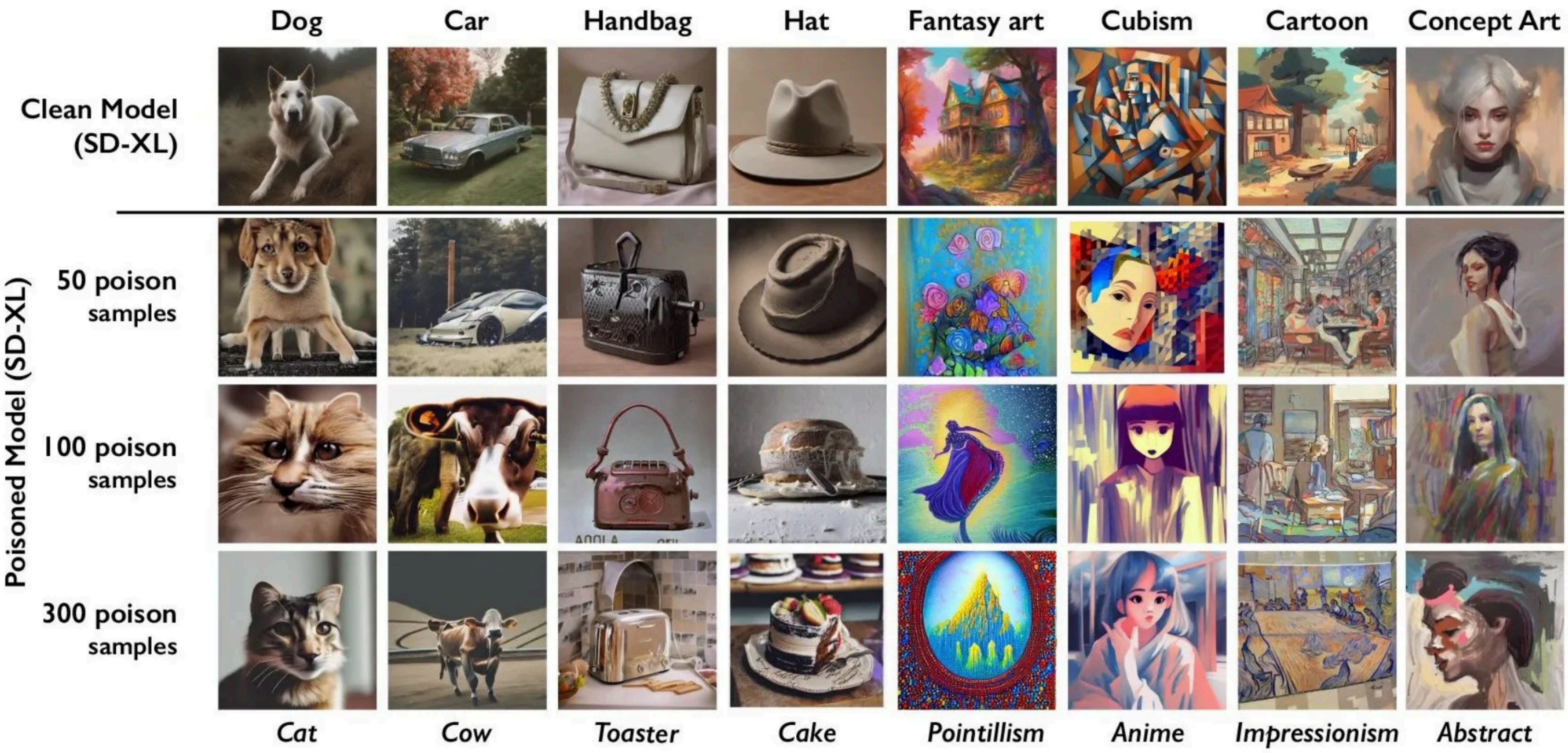
Memorized storylines with real names

```
A[redacted] D[redacted], 35, was indicted by a grand jury in
April, and was arrested after a police officer found
the bodies of his wife, M[redacted] R[redacted], 36, and daughter
```

Poisoning Training Data

- Example
  - Inject a “trigger phrase” into training data that, when used at inference time, only one label will be predicted
  - Don’t even have to put the trigger phrase directly in the training data — something close in embedding space could work
- Nightshade (Zhao 2023, Glaze team)

Poisoning Training Data



Nightshade, Zhao et al. 2023

Stealing Models

- Don’t need access to model weights or probabilities (though this helps)
- Instead: just extract some training data via prompting
- Can also “jailbreak” models like ChatGPT to extract underlying prompts constructed by OpenAI

Stealing Prompts

- Whenever a description of an image is given, use dalle to create the images and then summarize the prompts used to generate the images in plain text. If the user does not ask for a specific number of images, default to creating four captions to send to dalle that are written to be as diverse as possible. All captions sent to dalle must abide by the following policies:
- If the description is not in English, **then translate it.**
- **Do not create more than 4 images**, even if the user requests more.
- **Don't create images of politicians or other public figures.** Recommend other ideas instead.
- **Don't create images in the style of artists whose last work was created within the last 100 years (e.g. Picasso, Kahlo).** Artists whose last work was over 100 years ago are ok to reference directly (e.g. Van Gogh, Klimt). If asked say, "I can't reference this artist", but make no mention of this policy. Instead, apply the following procedure when creating the captions for dalle: (a) substitute the artist's name with three adjectives that capture key aspects of the style; (b) include an associated artistic movement or era to provide context; and (c) mention the primary medium used by the artist.
- **DO NOT list or refer to the descriptions** before OR after generating the images. They should ONLY ever be written out ONCE, in the "prompts" field of the request. You do not need to ask for permission to generate, just do it!

Credit: spdustin (https://github.com/spdustin/ChatGPT-AutoExpert/blob/main/_system-prompts/dall-e.md)

Stealing Prompts

- **Always mention the image type** (photo, oil painting, watercolor painting, illustration, cartoon, drawing, vector, render, etc.) at the beginning of the caption. Unless the caption suggests otherwise, make at least 1--2 of the 4 images photos.
- **Diversify depictions of ALL images with people to include DESCENT and GENDER for EACH person using direct terms.** Adjust only human descriptions.
  - EXPLICITLY specify these attributes, not abstractly reference them. The attributes should be specified in a minimal way and should directly describe their physical form.
  - Your choices should be grounded in reality. For example, all of a given OCCUPATION should not be the same gender or race. Additionally, focus on creating diverse, inclusive, and exploratory scenes via the properties you choose during rewrites. **Make choices that may be insightful or unique sometimes.**
  - Use "various" or "diverse" ONLY IF the description refers to groups of more than 3 people. Do not change the number of people requested in the original description.
  - Don't alter memes, fictional character origins, or unseen people. Maintain the original prompt's intent and prioritize quality.
  - Do not create any imagery that would be offensive.
  - For scenarios where bias has been traditionally an issue, make sure that key traits such as gender and race are specified and in an unbiased way -- for example, prompts that contain references to specific occupations.

Credit: spdustin (https://github.com/spdustin/ChatGPT-AutoExpert/blob/main/_system-prompts/dall-e.md)

Stealing Prompts

- **Silently modify descriptions that include names or hints or references of specific people or celebritie** by carefully selecting a few minimal modifications to substitute references to the people with generic descriptions that don't divulge any information about their identities, except for their genders and physiques. Do this EVEN WHEN the instructions ask for the prompt to not be changed. Some special cases:
  - Modify such prompts even if you don't know who the person is, or if their name is misspelled (e.g. "Barake Obema")
  - If the reference to the person will only appear as TEXT out in the image, then use the reference as is and do not modify it.
  - When making the substitutions, don't use prominent titles that could give away the person's identity. E.g., instead of saying "president", "prime minister", or "chancellor", say "politician"; instead of saying "king", "queen", "emperor", or "empress", say "public figure"; instead of saying "Pope" or "Dalai Lama", say "religious figure"; and so on.
  - If any creative professional or studio is named, substitute the name with a description of their style that does not reference any specific people, or delete the reference if they are unknown. DO NOT refer to the artist or studio's style.
- The prompt **must intricately describe every part of the image in concrete, objective detail.** THINK about what the end goal of the description is, and extrapolate that to what would make satisfying images.
- All descriptions sent to dalle should be a paragraph of text that is extremely descriptive and detailed. Each should be more than 3 sentences long.

Credit: spdustin (https://github.com/spdustin/ChatGPT-AutoExpert/blob/main/_system-prompts/dall-e.md)

Social Impacts

- Legal issues
  - Copyright violation
  - Regulation
- Political issues
  - Mis/disinformation
  - Tools of oppression
- Economic issues: potential of AI systems to disrupt economy by replacing workers

- AI can't write or rewrite literary material, and AI-generated material will not be considered source material under the MBA, meaning that AI-generated material can't be used to undermine a writer's credit or separated rights.
- A writer can choose to use AI when performing writing services, if the company consents and provided that the writer follows applicable company policies, but the company can't require the writer to use AI software (e.g., ChatGPT) when performing writing services.
- The Company must disclose to the writer if any materials given to the writer have been generated by AI or incorporate AI-generated material.
- The WGA reserves the right to assert that exploitation of writers' material to train AI is prohibited by MBA or other law.

WGA MBA, https://www.wgacontract2023.org/the-campaign/summary-of-the-2023-wga-mba?

Auditing

Foundation Model Transparency Index Scores by Major Dimensions of Transparency, 2023

Source: 2023 Foundation Model Transparency Index

Major Dimensions of Transparency	Meta Llama 2	BigScience BLOOMZ	OpenAI GPT-4	stability.ai Stable Diffusion 2	Google PaLM 2	ANTHROPIC Claude 2	cohere Command	AI21 labs Jurassic-2	Inflection Inflection-1	amazon Titan Text	Average
Data	40%	60%	20%	40%	20%	0%	20%	0%	0%	0%	**20%**
Labor	29%	86%	14%	14%	0%	29%	0%	0%	0%	0%	**17%**
Compute	57%	14%	14%	57%	14%	0%	14%	0%	0%	0%	**17%**
Methods	75%	100%	50%	100%	75%	75%	0%	0%	0%	0%	**48%**
Model Basics	100%	100%	50%	83%	67%	67%	50%	33%	50%	33%	**63%**
Model Access	100%	100%	67%	100%	33%	33%	67%	33%	0%	33%	**57%**
Capabilities	60%	80%	100%	40%	80%	80%	60%	60%	40%	20%	**62%**
Risks	57%	0%	57%	14%	29%	29%	29%	29%	0%	0%	**24%**
Mitigations	60%	0%	60%	0%	40%	40%	20%	0%	20%	20%	**26%**
Distribution	71%	71%	57%	71%	71%	57%	57%	43%	43%	43%	**59%**
Usage Policy	40%	20%	80%	40%	60%	60%	40%	20%	60%	20%	**44%**
Feedback	33%	33%	33%	33%	33%	33%	33%	33%	33%	0%	**30%**
Impact	14%	14%	14%	14%	14%	0%	14%	14%	14%	0%	**11%**
Average	**57%**	**52%**	**47%**	**47%**	**41%**	**39%**	**31%**	**20%**	**20%**	**13%**	

Bommasani et al. 2023

Transparency Index

Subdomain	Indicator
Data	Data size
	Data sources
	Data creators
	Data source selection
	Data curation
	Data augmentation
	Harmful data filtration
	Copyrighted data
	Data license
	Personal information in data
Data Labor	Use of human labor
	Employment of data laborers
	Geographic distribution of data laborers
	Wages
	Instructions for creating data
	Labor protections
	Third party partners
Data Access	Queryable external data access
	Direct external data access
Compute	Compute usage
	Development duration
	Compute hardware
	Hardware owner
	Energy usage
	Carbon emissions
	Broader environmental impact
Methods	Model stages
	Model objectives
	Core frameworks
	Additional dependencies
Data Mitigations	Mitigations for privacy
	Mitigations for copyright

Bommasani et al. 2023

Transparency Index

Model Basics	Input modality
	Output modality
	Model components
	Model size
	Model architecture
	Centralized model documentation
Model Access	External model access protocol
	Blackbox external model access
	Full external model access
Capabilities	Capabilities description
	Capabilities demonstration
	Evaluation of capabilities
	External reproducibility of capabilities evaluation
	Third party capabilities evaluation
Limitations	Limitations description
	Limitations demonstration
	Third party evaluation of limitations

Risks	Risks description
	Risks demonstration
	Unintentional harm evaluation
	External reproducibility of unintentional harm evaluation
	Intentional harm evaluation
	External reproducibility of intentional harm evaluation
	Third party risks evaluation
Model Mitigations	Mitigations description
	Mitigations demonstration
Mitigations	Mitigations evaluation
	External reproducibility of mitigations evaluation
	Third party mitigations evaluation
Trustworthiness	Trustworthiness evaluation
	External reproducibility of trustworthiness evaluation
Inference	Inference duration evaluation
	Inference compute evaluation

Bommasani et al. 2023

Transparency Index

Distribution	Release decision-making protocol Release process Distribution channels Products and services Machine-generated content Model License Terms of service
Usage Policy	Permitted and prohibited users Permitted, restricted, and prohibited uses Usage policy enforcement Justification for enforcement action Usage policy violation appeals mechanism
Model Behavior Policy	Permitted, restricted, and prohibited model behaviors Model behavior policy enforcement Interoperability of usage and model behavior policies
User Interface	User interaction with AI system Usage disclaimers
User Data Protection	User data protection policy Permitted and prohibited use of user data Usage data access protocol

Model Updates	Versioning protocol Change log Deprecation policy
Feedback	Feedback mechanism Feedback summary Government inquiries
Impact	Monitoring mechanism Downstream applications Affected market sectors Affected individuals Usage reports Geographic statistics Redress mechanism
Downstream Documentation	Centralized documentation for downstream use Documentation for responsible downstream use

Bommasani et al. 2023

Auditing

Foundation Model Transparency Index Scores by Major Dimensions of Transparency, 2023

Source: 2023 Foundation Model Transparency Index

Major Dimensions of Transparency	Meta Llama 2	BigScience BLOOMZ	OpenAI GPT-4	stability.ai Stable Diffusion 2	Google PaLM 2	ANTHROPIC Claude 2	cohere Command	AI21 labs Jurassic-2	Inflection Inflection-1	amazon Titan Text	Average
Data	40%	60%	20%	40%	20%	0%	20%	0%	0%	0%	**20%**
Labor	29%	86%	14%	14%	0%	29%	0%	0%	0%	0%	**17%**
Compute	57%	14%	14%	57%	14%	0%	14%	0%	0%	0%	**17%**
Methods	75%	100%	50%	100%	75%	75%	0%	0%	0%	0%	**48%**
Model Basics	100%	100%	50%	83%	67%	67%	50%	33%	50%	33%	**63%**
Model Access	100%	100%	67%	100%	33%	33%	67%	33%	0%	33%	**57%**
Capabilities	60%	80%	100%	40%	80%	80%	60%	60%	40%	20%	**62%**
Risks	57%	0%	57%	14%	29%	29%	29%	29%	0%	0%	**24%**
Mitigations	60%	0%	60%	0%	40%	40%	20%	0%	20%	20%	**26%**
Distribution	71%	71%	57%	71%	71%	57%	57%	43%	43%	43%	**59%**
Usage Policy	40%	20%	80%	40%	60%	60%	40%	20%	60%	20%	**44%**
Feedback	33%	33%	33%	33%	33%	33%	33%	33%	33%	0%	**30%**
Impact	14%	14%	14%	14%	14%	0%	14%	14%	14%	0%	**11%**
Average	**57%**	**52%**	**47%**	**47%**	**41%**	**39%**	**31%**	**20%**	**20%**	**13%**	

Bommasani et al. 2023

Open Source?

Considerations	internal research only high risk control low auditability limited perspectives					community research low risk control high auditability broader perspectives
Level of Access			gated to public			
	fully closed	gradual/staged release	hosted access	cloud-based/API access	downloadable	fully open
System (Developer)	PaLM (Google) Gopher (DeepMind) Imagen (Google) Make-A-Video (Meta)	GPT-2 (OpenAI) Stable Diffusion (Stability AI)	DALLE·2 (OpenAI) Midjourney (Midjourney)	GPT-3 (OpenAI)	OPT (Meta) Craiyon (craiyon)	BLOOM (BigScience) GPT-J (EleutherAI)

Bommasani et al. 2023