Advanced Natural Language Processing (ANLP)

Lecture 7: Spurious Correlations in NLP Benchmarks

Gabriel Stanovsky & **Roy Schwartz**

THE HEBREW UNIVERSITY OF JERUSALEM

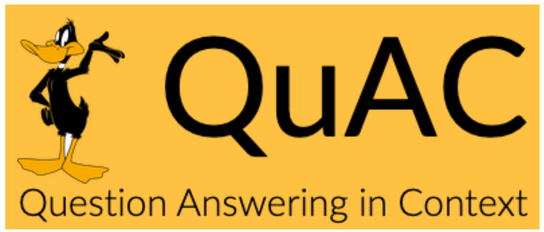
Benchmarks in NLP

Natural Questions

A Benchmark for Question Answering Research.

SWAG

DROP



MultiNLI

SuperGLUE

SQuAD2.0
The Stanford Question Answering Dataset

Benchmarks in NLP

The Premise

SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference

Rowan Zellers[♠] **Yonatan Bisk**[♠] **Roy Schwartz**[♠♡] **Yejin Choi**[♠♡]
[♠]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♡]Allen Institute for Artificial Intelligence
{rowanz,ybisk,roysch,yejin}@cs.washington.edu
https://rowanzellers.com/swag

2017). **First**, our dataset poses a new challenge of grounded commonsense inference that is easy for humans (88%) while hard for current state-of-the-art NLI models (<60%). **Second**, our pro-

Benchmarks in NLP

The Reality

Benchmark	Baseline	Shortly after
SWAG (Zellers, Bisk, Schwartz & Choi, 2018)	52%	86% (Devlin et al., 2018)
DROP (Dua, Wang, Dasigi, Stanovsky et al., 2019)	47 F1	90 F1 (Chen et al., 2020)
HellaSWAG (Zellers et al., 2019)	47%	93% (He et al., 2020)
WinoGrande (Sakaguchi et al., 2020)	53% AUC	88% AUC (Raffel et al., 2020)

Spurious Correlations

*In statistics, a spurious relationship or spurious correlation is a mathematical relationship in which **two or more events or variables** are **associated** but not **causally related**, due to either **coincidence** or the presence of a **certain third, unseen factor**. Wikipedia*

Spurious Correlations and NLP Benchmarks

- Instead of **understanding** the text, machines pick up on these **correlations** from the training data
  - They use the learned correlations to excel on the test sets
- This artificially **inflate** the **state of the art**
- As a result, many efforts exist to **mitigate these correlations**

Outline

- Spurious correlations in NLP
- Mitigating spurious correlations
  - Adversarial networks: Modify the model
  - Challenge sets: Modify the test data
  - Balancing/filtering: Modify the training data
- Revisiting spurious correlations

Visual Question Answering

VQA Dataset (Antol et al., 2015)

- Input: an image and a question
  - What sport is this man playing?
  - Do you see a shadow?
- Output: answer
  - Tennis, yes

Spurious Correlations in VQA

Zhang et al. (2016); Goyal et al. (2017)

- 40% of the questions in VQA starting with "***What sport is this***" are answered with "***tennis***"
- "***yes***" is the answer to 87% of the questions in the VQA dataset starting with "***Do you see a***"

ROC Story Cloze Task

Mostafazadeh et al. (2016)

Context	Right Ending	Wrong Ending
Tom and Sheryl have been together for two years. One day, they went to a carnival together. He won her several stuffed bears, and bought her funnel cakes. When they reached the Ferris wheel, he got down on one knee.	Tom asked Sheryl to marry him.	He wiped mud off of his boot.

- A story comprehension task
- The task: given a story prefix, distinguish between the coherent and the incoherent endings

The Effect of Writing Style

Schwartz et al (2017); Cai et al. (2017)

- Train a binary classifier on **the endings only**
  - Ignoring the story prefix

Right Ending	Wrong Ending
Tom asked Sheryl to marry him.	He wiped mud off of his boot.

Right	Weight	Freq.	*Wrong*	Weight	Freq.
'ed .'	0.17	6.5%	START NNP	0.21	54.8%
'and '	0.15	13.6%	NN .	0.17	47.5%
JJ	0.14	45.8%	NN NN .	0.15	5.1%
to VB	0.13	20.1%	VBG	0.11	10.1%
'd th'	0.12	10.9%	START NNP VBD	0.11	41.9%

Model	**Acc.**
DSSM (Mostafazadeh et al., 2016a)	0.585
ukp (Bugert et al., 2017)	0.717
tbmihaylov (Mihaylov and Frank, 2017)	0.724
†EndingsOnly (Cai et al., 2017)	0.725
cogcomp	0.744
HIER,ENCPLOTEND,ATT (Cai et al., 2017)	0.747
RNN	0.677
†Ours	0.724
Combined (ours + RNN)	**0.752**
Human judgment	1.000

Natural Language Inference (NLI)

SNLI (Bowman et al., 2015); MNLI (Williams et al., 2018)

Premise	A woman selling bamboo sticks talking to two men on a loading dock.
Entailment	There are **at least** three **people** on a loading dock.
Neutral	A woman is selling bamboo sticks **to help provide for her family.**
Contradiction	A woman is **not** taking money for any of her sticks.

Annotation Artifacts in NLI Datasets

Gururangan, Swayamdipta, Levy, Schwartz et al. (2018); Poliak et al. (2018); Tsuchiya (2018)

- Train a hypothesis-only classifier
  - No premise

	Entailment		Neutral		Contradiction	
SNLI	outdoors	2.8%	tall	0.7%	nobody	0.1%
	least	0.2%	first	0.6%	sleeping	3.2%
	instrument	0.5%	competition	0.7%	no	1.2%
	outside	8.0%	sad	0.5%	tv	0.4%
	animal	0.7%	favorite	0.4%	cat	1.3%
MNLI	some	1.6%	also	1.4%	never	5.0%
	yes	0.1%	because	4.1%	no	7.6%
	something	0.9%	popular	0.7%	nothing	1.4%
	sometimes	0.2%	many	2.2%	any	4.1%
	various	0.1%	most	1.8%	none	0.1%

Model	SNLI	MultiNLI	
		Matched	Mismatched
majority class	34.3	35.4	35.2
`fastText`	**67.0**	**53.9**	**52.3**

Other Spurious Correlations

- Question answering
  - Kaushik & Lipton (2018)
  - Sen & Saffari (2020)
- Common-sense reasoning
  - Elazar et al. (2021)
- Are We Modeling the Task or the Annotator?
  - Geva et al. (2019)

Outline

- Spurious correlations in NLP
- Mitigating spurious correlations
  - Adversarial networks: Modify the model
  - Challenge sets: Modify the test data
  - Balancing/filtering: Modify the training data
- Revisiting spurious correlations

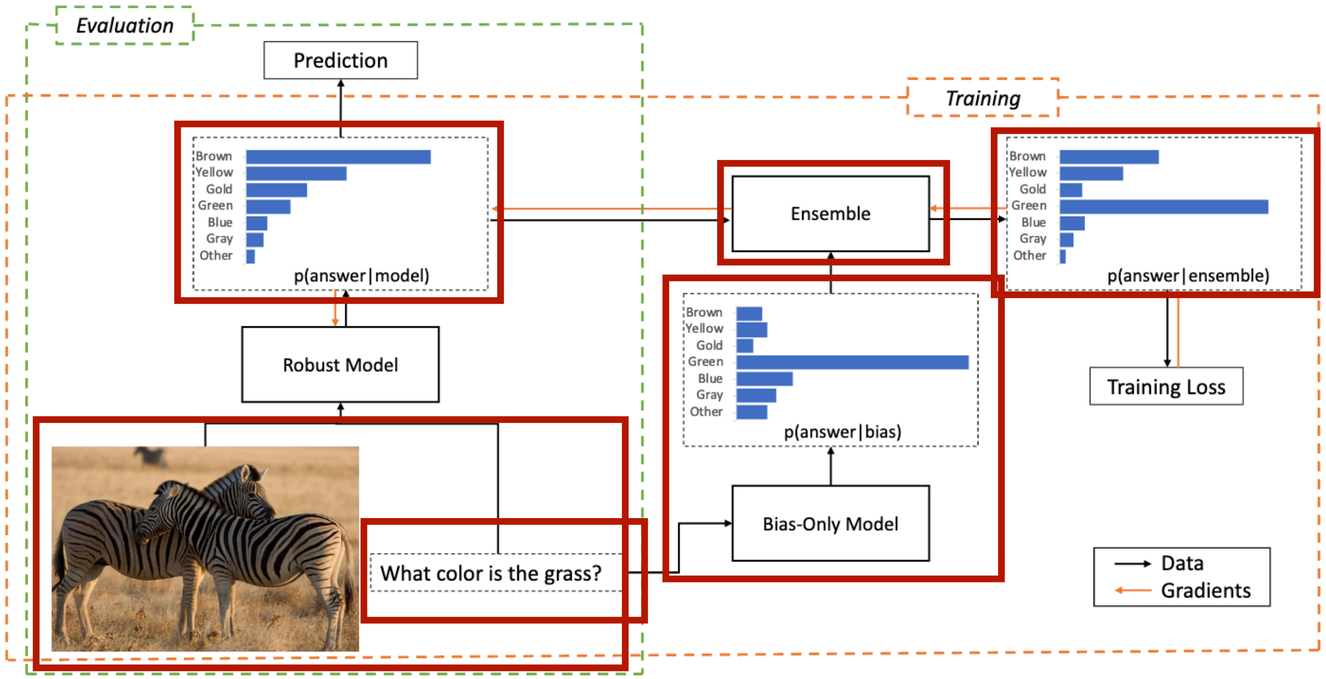
Mitigating Spurious Correlations

Adversarial Networks: Modify the Model

Model Ensemble

Clark et al. (2019)

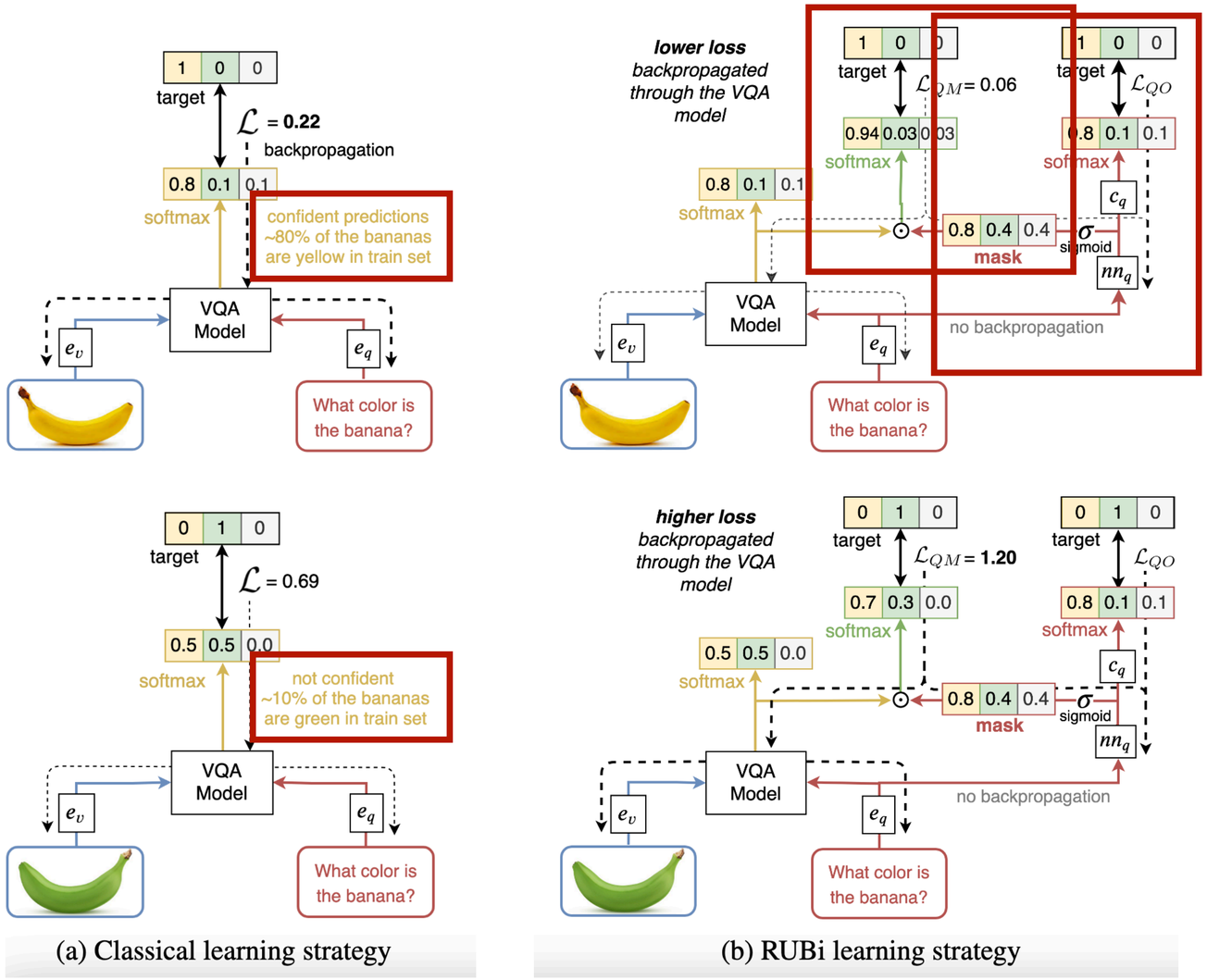
- Start by training a partial input model
  - Which tends to pick up on spurious features
- Then train a full model in an ensemble with the first model
- The robust model has less incentive to learn spurious features, and thus generalizes better



Loss Modification

Cadene et al. (2019)

- Train a partial input model
- Use its predictions to modify the loss of the main model
  - Loss on biased examples decreases, loss on non-biased examples increases



(a) Classical learning strategy

(b) RUBi learning strategy

Mitigating Spurious Correlations

Challenge Sets: Modify the Test Data

Challenge Sets

- NLP models are very sensitive to their training domain
- Testing a model on a different distribution often leads to reduced performance
  - Fixing this problem is one of the key challenges in NLP and AI in general
- Challenge dataset (aka *adversarial datasets*) intentionally aim to mislead the model
  - The goal is to uncover specific model weaknesses

Adversarial SQuAD

Jia et al. (2017)

SQuAD1.1 Leaderboard (Rajpurkar et al., 2016)

Here are the ExactMatch (EM) and F1 scores evaluated on the test set of SQuAD v1.1.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Apr 10, 2020	LUKE (single model) *Studio Ousia & NAIST & RIKEN AIP*	**90.202**	**95.379**
2 May 21, 2019	XLNet (single model) *Google Brain & CMU*	89.898	95.080
3 Dec 11, 2019	XLNET-123++ (single model) *MST/EOI* http://tia.today	89.856	94.903

Article: Super Bowl 50

Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager.*

Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"

NLI Stress Test

Naik et al. (2018)

Error Cat.	Premise	Hypothesis
Antonyms	I love the Cinderella story.	I hate the Cinderella story.
Numerical Reasoning	Tim has 350 pounds of cement in 100, 50, and 25 pound bags	Tim has less than 750 pounds of cement in 100, 50, and 25 pound bags
Word Overlap	Possibly no other country has had such a turbulent history.	The country's history has been turbulent and true is true
Negation	Possibly no other country has had such a turbulent history.	The country's history has been turbulent and false is not true
Length Mismatch	Possibly no other country has had such a turbulent history and true is true and true is true and true is true and true is true and true is true	The country's history has been turbulent.
Spelling Errors	As he emerged, Boris remarked, glancing up at teh clock: "You are early	Boris had just arrived at the rendezvous when he appeared

System	Original MultiNLI Dev		Competence Test			Distraction Test						Noise Test	
			Antonymy		Numerical Reasoning	Word Overlap		Negation		Length Mismatch		Spelling Error	
	Mat	Mis	Mat	Mis		Mat	Mis	Mat	Mis	Mat	Mis	Mat	Mis
NB	74.2	74.8	15.1	19.3	21.2	47.2	47.1	39.5	40.0	48.2	47.3	51.1	49.8
CH	73.7	72.8	11.6	9.3	30.3	58.3	58.4	52.4	52.2	63.7	65.0	68.3	69.1
RC	71.3	71.6	36.4	32.8	30.2	53.7	54.4	49.5	50.4	48.6	49.6	66.6	67.0
IS	70.3	70.6	14.4	10.2	28.8	50.0	50.2	46.8	46.6	58.7	59.4	58.3	59.4
BiLSTM	70.2	70.8	13.2	9.8	31.3	57.0	58.5	51.4	51.9	49.7	51.2	65.0	65.1
CBOW	63.5	64.2	6.3	3.6	30.3	53.6	55.6	43.7	44.2	48.0	49.3	60.3	60.6

Challenge Sets

- Test various Types of Capabilities
  - Shift in distribution
  - Ignoring noise
  - Handling misspellings
  - Handling negation
  - Handling temporal modifications
- Applied to a Range of NLP Tasks
  - NLI
  - (Visual-/)Question answering
  - Machine Translation
  - Text classification
  - …

CheckList for NLP Models

Ribeiro et al. (2020)

- Treating NLP models as programs
- Evaluating them using **unit tests**

Labels: positive, negative, or neutral; INV: same pred. (INV) after removals/ additions; DIR: sentiment should not decrease (↑) or increase (↓)

	Test *TYPE* and Description	Failure Rate (%)					Example test cases & expected behavior
						RoB	
Vocab.+POS	***MFT:*** Short sentences with neutral adjectives and nouns	0.0	7.6	4.8	94.6	81.8	The company is Australian. neutral That is a private aircraft. neutral
	MFT: Short sentences with sentiment-laden adjectives	4.0	15.0	2.8	0.0	0.2	That cabin crew is extraordinary. pos I despised that aircraft. neg
	INV: Replace neutral words with other neutral words	9.4	16.2	12.4	10.2	10.2	@Virgin should I be concerned that → when I'm about to fly ... INV @united the → our nightmare continues... INV
	DIR: Add positive phrases, fails if sent. goes down by > 0.1	12.6	12.4	1.4	0.2	10.2	@SouthwestAir Great trip on 2672 yesterday... You are extraordinary. ↑ @AmericanAir AA45 ... JFK to LAS. You are brilliant. ↑
	DIR: Add negative phrases, fails if sent. goes up by > 0.1	0.8	34.6	5.0	0.0	13.2	@USAirways your service sucks. You are lame. ↓ @JetBlue all day. I abhor you. ↓
Robust.	***INV:*** Add randomly generated URLs and handles to tweets	9.6	13.4	24.8	11.4	7.4	@JetBlue that selfie was extreme. @pi9QDK INV @united stuck because staff took a break? Not happy 1K.... https://t.co/PWK1jb INV
	INV: Swap one character with its neighbor (typo)	5.6	10.2	10.4	5.2	3.8	@JetBlue → @JeBtlue I cri INV @SouthwestAir no thanks → thakns INV
NER	***INV:*** Switching locations should not change predictions	7.0	20.8	14.8	7.6	6.4	@JetBlue I want you guys to be the first to fly to # Cuba → Canada... INV @VirginAmerica I miss the #nerdbird in San Jose → Denver INV
	INV: Switching person names should not change predictions	2.4	15.1	9.1	6.6	2.4	...Airport agents were horrendous. Sharon → Erin was your saviour INV @united 8602947, Jon → Sean at http://t.co/58tuTgli0D, thanks. INV
Temporal	***MFT:*** Sentiment change over time, present should prevail	41.0	36.6	42.2	18.8	11.0	I used to hate this airline, although now I like it. pos In the past I thought this airline was perfect, now I think it is creepy. neg
Negation	***MFT:*** Negated negative should be positive or neutral	18.8	54.2	29.4	13.2	2.6	The food is not poor. pos or neutral It isn't a lousy customer service. pos or neutral
	MFT: Negated neutral should still be neutral	40.4	39.6	74.2	98.4	95.4	This aircraft is not private. neutral This is not an international flight. neutral
	MFT: Negation of negative at the end, should be pos. or neut.	100.0	90.4	100.0	84.8	7.2	I thought the plane would be awful, but it wasn't. pos or neutral I thought I would dislike that plane, but I didn't. pos or neutral
	MFT: Negated positive with neutral content in the middle	98.4	100.0	100.0	74.0	30.2	I wouldn't say, given it's a Tuesday, that this pilot was great. neg I don't think, given my history with airplanes, that this is an amazing staff. neg
SRL	***MFT:*** Author sentiment is more important than of others	45.4	62.4	68.0	38.8	30.0	Some people think you are excellent, but I think you are nasty. neg Some people hate you, but I think you are exceptional. pos
	MFT: Parsing sentiment in (question, "yes") form	9.0	57.6	20.8	3.6	3.0	Do I think that airline was exceptional? Yes. neg Do I think that is an awkward customer service? Yes. neg
	MFT: Parsing sentiment in (question, "no") form	96.8	90.8	81.6	55.4	54.8	Do I think the pilot was fantastic? No. neg Do I think this company is bad? No. pos or neutral

Table 1: A selection of tests for sentiment analysis. All examples (right) are failures of at least one model.

Contrast Sets

Gardner et al. (2020)

- Manually perturb the test instances in small but meaningful ways that (typically) change the gold label
- Typically generated manually
  - Although see Bitton, Stanovsky, Schwartz & Elhadad (2021)

Original Example:

Two similarly-colored and similarly-posed chow dogs are face to face in one image.

Mitigating Spurious Correlations

Balancing/Filtering: Modify the Training Data

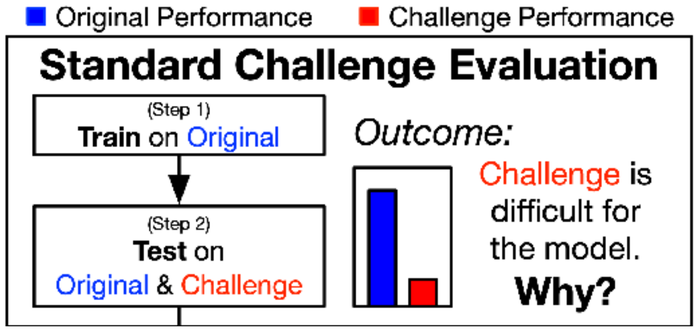
Dataset Balancing

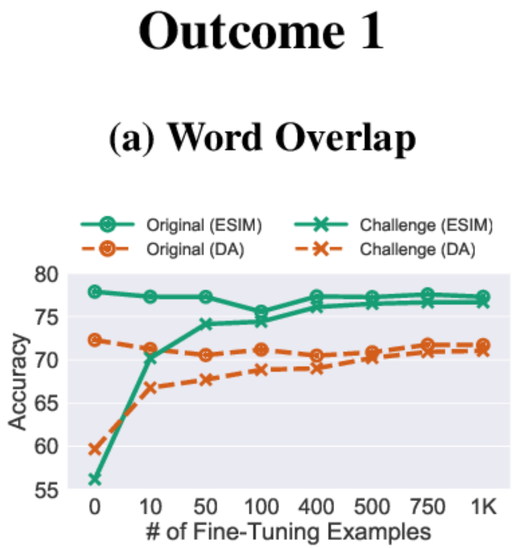
Augmentation

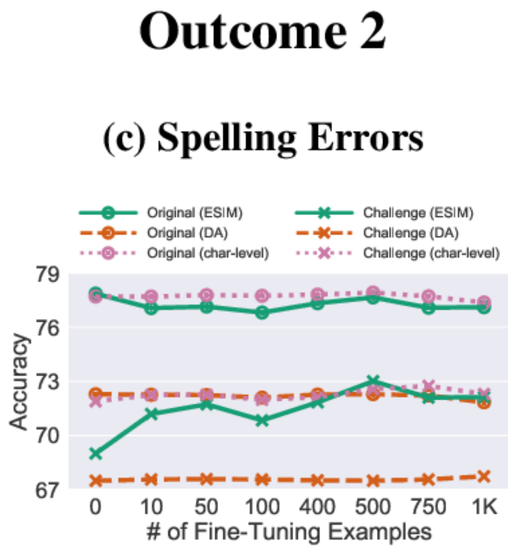
- The key idea: balance-out spurious correlations
  - Similar idea to contrast sets, applied to training sets
- Vision and Language datasets
  - VQA 2.0 (Goyal et al., 2017)
  - GQA (Hudson and Manning, 2019)
- Language only
  - ROC stories cloze task 1.5 (Sharma et al., 2018)

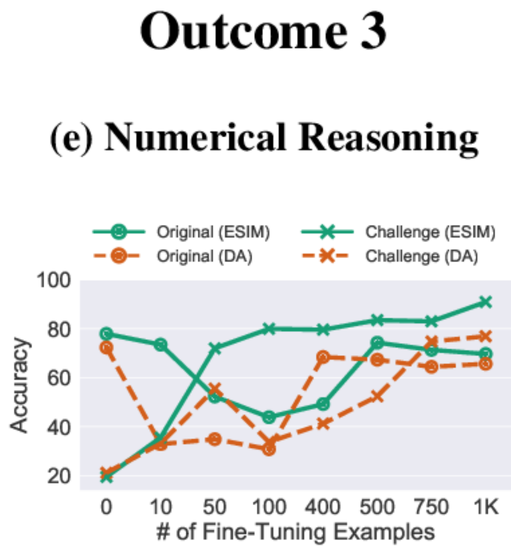
Inoculation by Fine-Tuning 💉

Liu, Schwartz, Smith (2019)





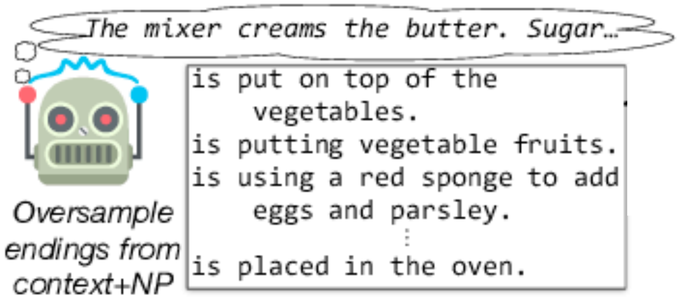




Adversarial Filtering (AF)

Zellers, Bisk, Schwartz & Choi (2018)

- A multi-choice setting
- An LM generates *many* possible distractors
- A discriminator trained to identify the machine-generated options
- Iteratively until convergence:
  - Select easily-identifiable options
  - Replace them with other (harder) options
- Validate resulting data with human experts



Adversarial Filters of Dataset Biases (AFLite)

Sakaguchi et al. (2020)

- Start from a collected dataset *D*
- Iteratively
  - Randomly break *D* into *n* different train/test splits
  - Train a classifier on each training split
  - Filter out the instances that are solved by most models
- Return filtered dataset

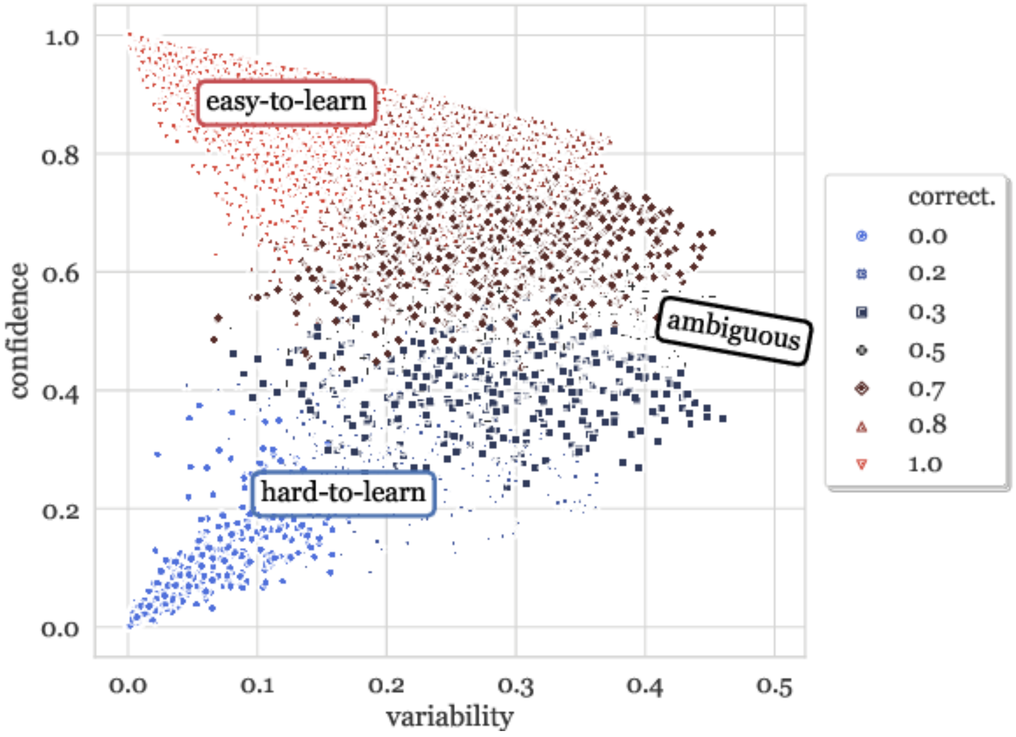
Filtering as Balancing

- As the adversarial model grows, models will pick up subtler correlations
  - Resulting in a fully *balanced* dataset
- Widely adopted
  - Record (Zhang et al., 2018)
  - DROP (Dua et al., 2019)
  - HellaSWAG (Zellers et al., 2019)
  - αNLI (Bhagavatula et al., 2019)
  - WinoGrande (Sakaguchi et al., 2020)
  - …

Dataset Cartography

Swayamdipta, Schwartz et al. (2020)

- Identify different regions in datasets
- Most examples are *easy-to-learn*
- Training on the most ambiguous examples leads to **better generalization**

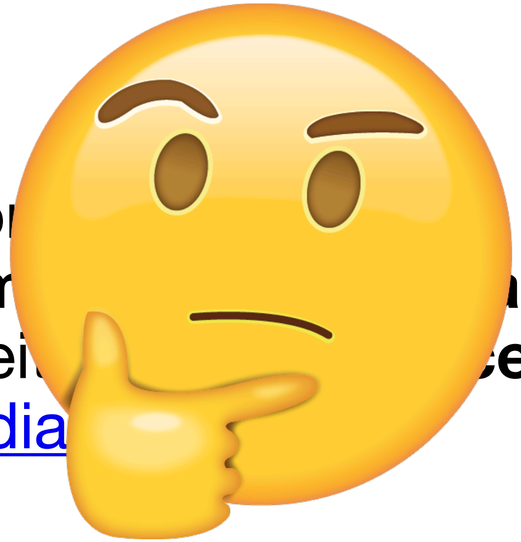


Outline

- Spurious correlations in NLP
- Mitigating spurious correlations
  - Adversarial networks: Modify the model
  - Challenge sets: Modify the test data
  - Balancing/filtering: Modify the training data
- Revisiting spurious correlations

What are Spurious Correlations?

- In statistics, a spurious relatio correlation is a mathematical relationship in which **two or m riables** are **associated** but not **causally related**, due to ei ce or the presence of a **certain third, unseen factor**. Wikipedia



Ingenuine correlations

- A feature correlated with some output label for no apparent reason
  - E.g., “cat” and “sleeping” are correlated with contradictions in SNLI (Gururangan, Swayamdipta, Levy, Schwartz et al., 2018)
  - Wang and Culotta, 2020; Rogers, 2021
- An appealing definition
- But somewhat subjective
  - E.g., the word “not” indicating NLI contradictions; “amazing” as a feature for positive sentiment

Ungeneralizable correlations

- A feature that works well for specific examples but **does not hold in general**
  - Chang et al., 2021; Yaghoobzadeh et al., 2021
- Does not address the nature of the feature
  - Whether genuine or not
- But does assume the feature is *important*
  - And thus somewhat subjective

every-word

- *Every* simple correlation between single word features and output labels is spurious
  - Gardner et al. (2021)
- *Competent* datasets: the marginal probability for every feature is uniform over the class label
  - $\forall x_i, y \in Y, p(y \mid x_i) = \frac{1}{|Y|}$

The Lost Battle Against Spurious Correlations?

Schwartz & Stanovsky (2022)

- Is balancing the right way forward?
  - Balancing too little is insufficient, balancing too much leaves nothing to learn
- Balancing is not even necessarily desired
  - May prevent us from learning world knowledge and common-sense knowledge

Split	Text	Label
Train	very good	+
	very bad	−
	not good	−
	not bad	+

Who is the president of the U.S.?

Context	Answer
∅	Joe Biden
The year 2019	Donald Trump
The West Wing, season 1	Josiah “Jed” Bartlet

Suggested Alternatives

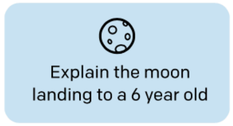
- Instead of balancing, augment datasets with *richer contexts*
- Instead of a closed label set, support *abstention/interaction*
- Instead of large-scale fine-tuning, move to *few-shot learning*

Reminder: InstructGPT Illustrated

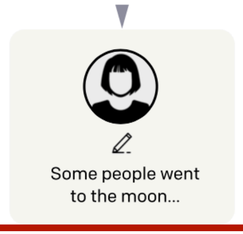
Step 1

Collect demonstration data, and train a supervised policy.

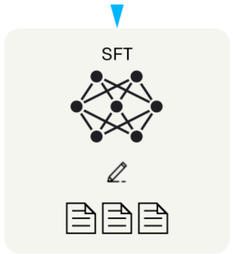
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



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

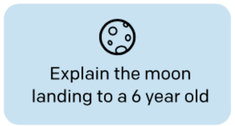


Instruction-Tuning

Step 2

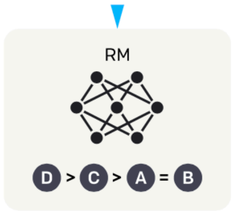
Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

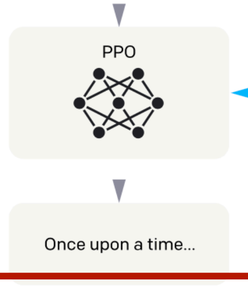


Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.



The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

RLHF

RLHF vs. Instruction-Tuning

- RL is harder than fine-tuning
- A weaker signal
  - One score per generated **text** vs. one score per generated **word**
- Requires a scoring function
  - Typically a human, which is **slow** and **expensive**

Why RLHF?
Diversity

- Fine-tuning rewards a single correct answer
- RL allows to give positive scores to many “correct” answers

https://gist.github.com/yoavg/6bff0fecd65950898eba1bb321cfbd81

Why RLHF?

Spurious Correlations

- A supervised model is incentivized to always give some answer, even if it doesn't know the answer
- This could lead to over-reliance on **spurious correlations**
- An RL model can get partial credit if it abstains rather than predicting the wrong answer

https://gist.github.com/yoavg/6bff0fecd65950898eba1bb321cfbd81

Example

What is the capital of France?

Model Answer	RL	Instruction-Tuning
Paris	Positive reward	0 loss
London	Negative reward	High loss
I don’t know	Medium reward	High loss

RLHF

Open Questions

- Improve supervised setup?
  - Make models abstain
  - Accept diverse answers
- Can we build testable hypotheses as to why RLHF works?

Fight Bias with Bias

Reif & Schwartz (2023)

- Balancing only hides the problem
  - Some biases remain hidden in the data
- We want models that are robust to such biases
- Let's ***amplify*** the biases in the data

Amplify Biases???

- Models are sensitive even to very subtle biases, which are hard to detect and filter

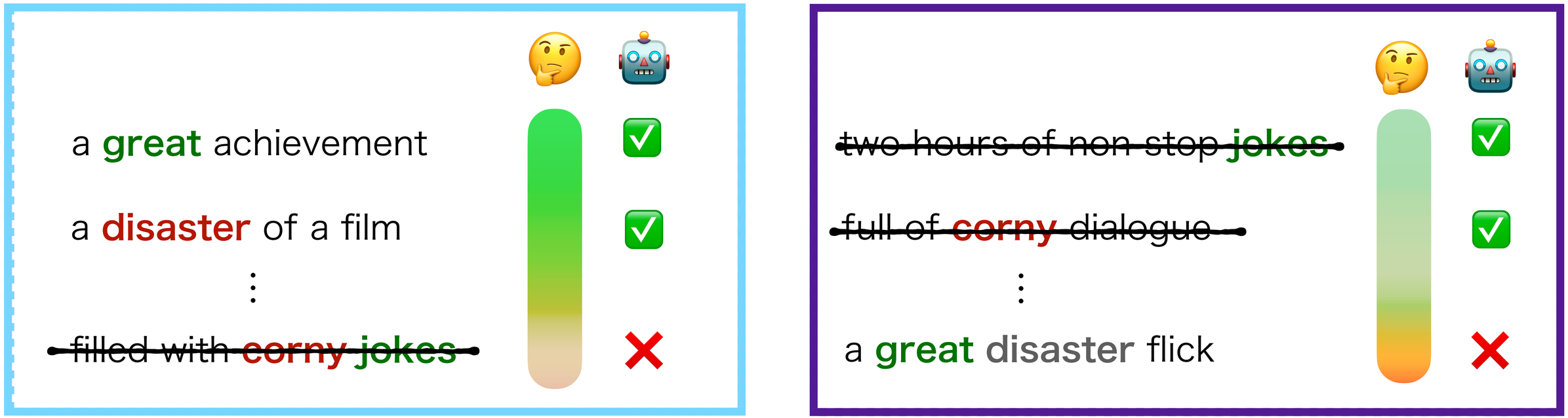
Biases "hide" in hard, filtered training sets

⇒ Harder to evaluate impact on models

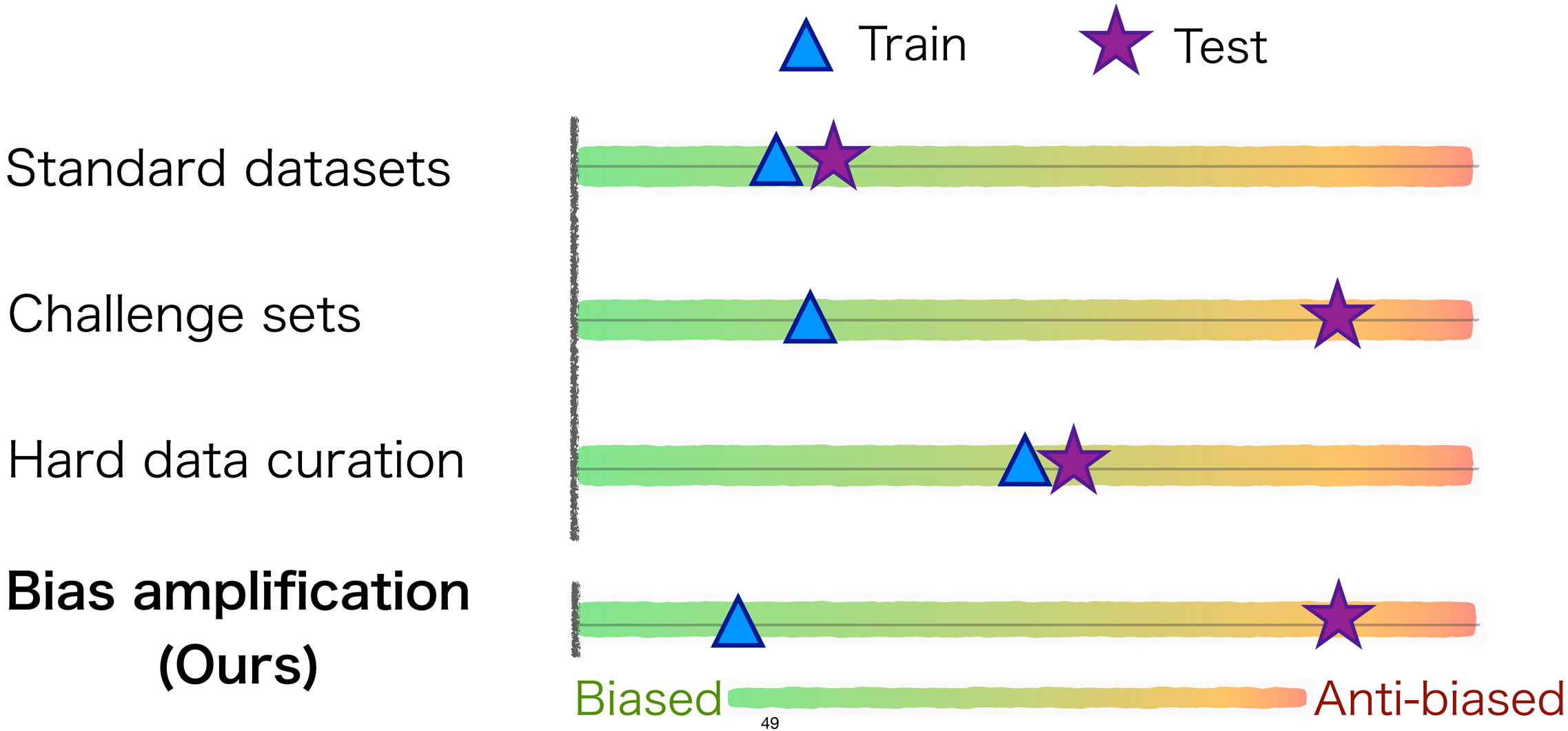
- Could we ever create datasets that don't contain exploitable biases?
  - Linzen et al. (2020); Schwartz & Stanovsky (2022)

Don’t Filter, Amplify

Bias-amplified Splits: *Biased* Training, *Anti-biased* Test

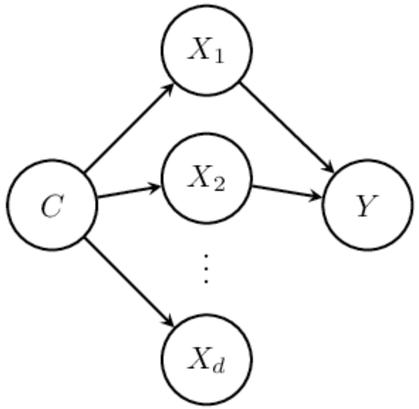


Model Evaluation Under Biases



An Analysis through a Causal Lens

Feder et al. (2022); Joshi et al. (2022)



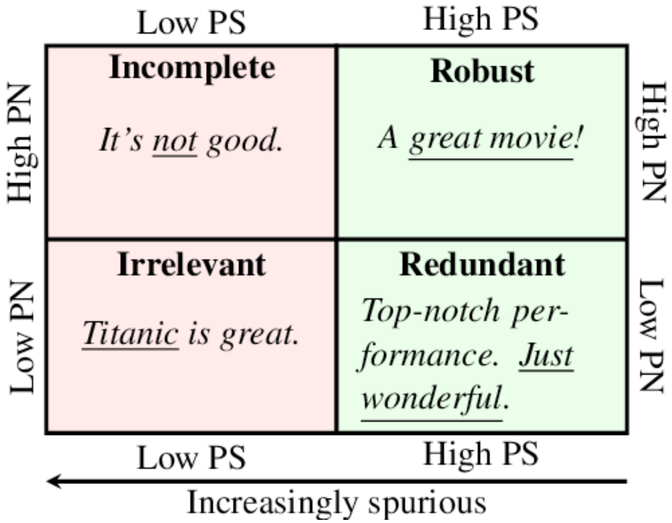
(a) Data generating model.

- C - cause
- X_i - words
- Y - label (sentiment)

Necessity and Sufficiency

Joshi et al. (2022)

- Study two characteristics of features
  - (N)ecessity
  - (S)ufficiency
- Define probabilities for each property
  - PN, PS
- Debiasing affects both types differently
  - Low PS features can be debiased, high PS cannot
  - Removing high PS features hurts performance



Summary

- Spurious correlations in NLP
- Mitigating spurious correlations
  - Adversarial networks: Modify the model
  - Challenge sets: Modify the test data
  - Balancing/filtering: Modify the training data
- Revisiting spurious correlations