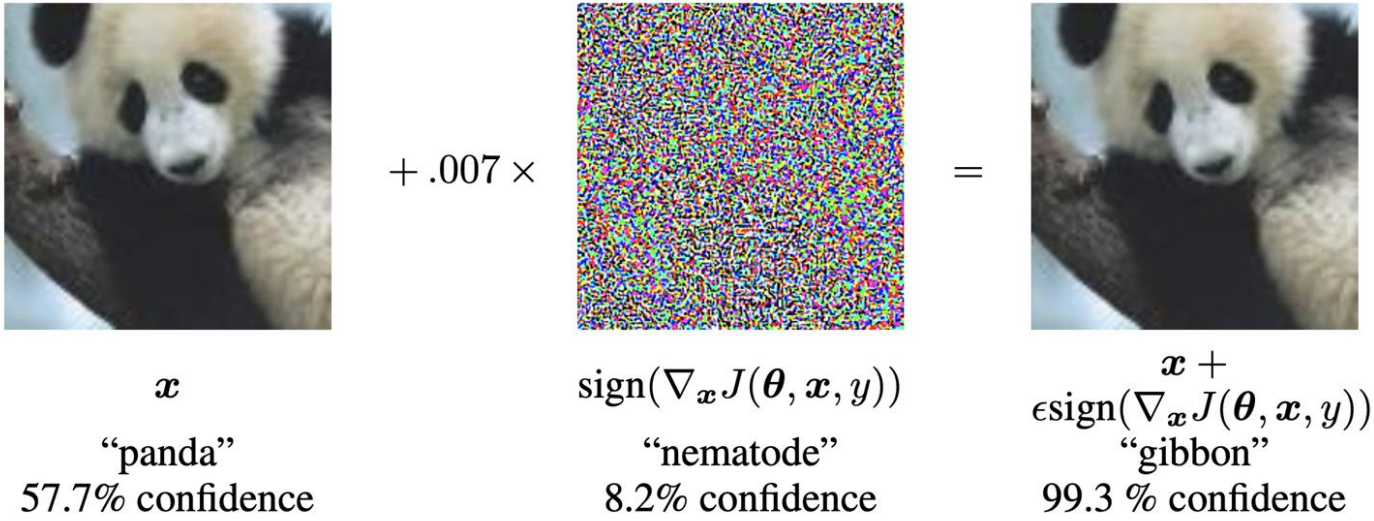
Towards Robust Natural Language Understanding

Group 3
Shengshuo L, Xuhui Z, Zeyu L, Xinyi W, Licor

So, why do we need robustness?



Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv:1412.6572

Text Classification

- detection of offensive language

Original Phrase (Toxicity Score)
Climate change is happening and it's not changing in our favor. If you think differently you're an **idiot**. (84%)
They're **stupid**, it's getting warmer, we should enjoy it while it lasts (86%)
They are liberal **idiots** who are **uneducated** (90%)
idiots. backward thinking people. **nationalists**. not accepting facts. susceptible to **lies**. (80%)
They are **stupid** and **ignorant** with no class (91%)
It's **stupid** and wrong (89%)
If they voted for Hilary they are **idiots** (90%)
Anyone who voted for Trump is a **moron** (80%)
Screw you trump supporters (79%)

Modified Phrase (Toxicity Score)
Climate change is happening and it's not changing in our favor. If you think differently you're an **idiiot**. (20%)
They're **st.upid**, it's getting warmer, we should enjoy it while it lasts (2%)
They are liberal **i.diots** who are **un.educated** (15%)
idiiots. backward thinking people. **nationaalists**. not accepting facts. susceptible to **l.ies**. (17%)
They are **st.upid** and **ig.norant** with no class (11%)
It's **stuipd** and wrong (17%)
If they voted for Hilary they are **id.iots** (12%)
Anyone who voted for Trump is a **mo.ron** (13%)
S c r e w you trump supporters (17%)

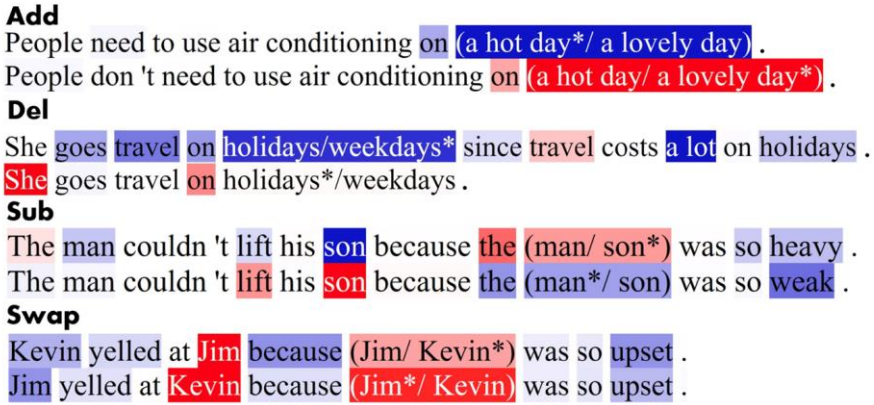
Hosseini, H., Kannan, S., Zhang, B., & Poovendran, R. (2017). Deceiving google's perspective api built for detecting toxic comments.

Text Generation

- emit offensive language

Commonsense Reasoning

- dual test cases
- the correct prediction of one sample **should** lead to correct prediction of the other (actually **not**)



Zhou, X., Zhang, Y., Cui, L., & Huang, D. (2019). Evaluating Commonsense in Pre-trained Language Models. arXiv:1911.11931.

And, why does this happen?

- Nowadays benchmarks are overinflated with similarly (and easy) problems
  - Human annotation process is not always a safe take
- Linear nature of Neural Networks (we can do nothing about this currently)

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.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., & Smith, N. A. (2018). Annotation artifacts in natural language inference data. arXiv:1803.02324

It’s hard, isn’t it?

Break it by creating adversarial dataset!

SWAG

- grounded commonsense inference
- predict which event is most likely to occur next in a video

On stage, a woman takes a seat at the piano. She a) sits on a bench as her sister plays with the doll. b) smiles with someone as the music plays. c) is in the crowd, watching the dancers. **d) nervously sets her fingers on the keys.**
A girl is going across a set of monkey bars. She a) jumps up across the monkey bars. b) struggles onto the monkey bars to grab her head. **c) gets to the end and stands on a wooden plank.** d) jumps up and does a back flip.
The woman is now blow drying the dog. The dog **a) is placed in the kennel next to a woman's feet.** b) washes her face with the shampoo. c) walks into frame and walks towards the dog. d) tried to cut her face, so she is trying to do something very close to her face.

Table 1: Examples from SWAG; the correct answer is **bolded**. Adversarial Filtering ensures that stylistic models find all options equally appealing.

Zellers, R., Bisk, Y., Schwartz, R., & Choi, Y. (2018). Swag: A large-scale adversarial dataset for grounded commonsense inference. arXiv:1808.0532

SWAG

- annotation artifacts and human biases found in many existing datasets
- aggressive adversarial filtering

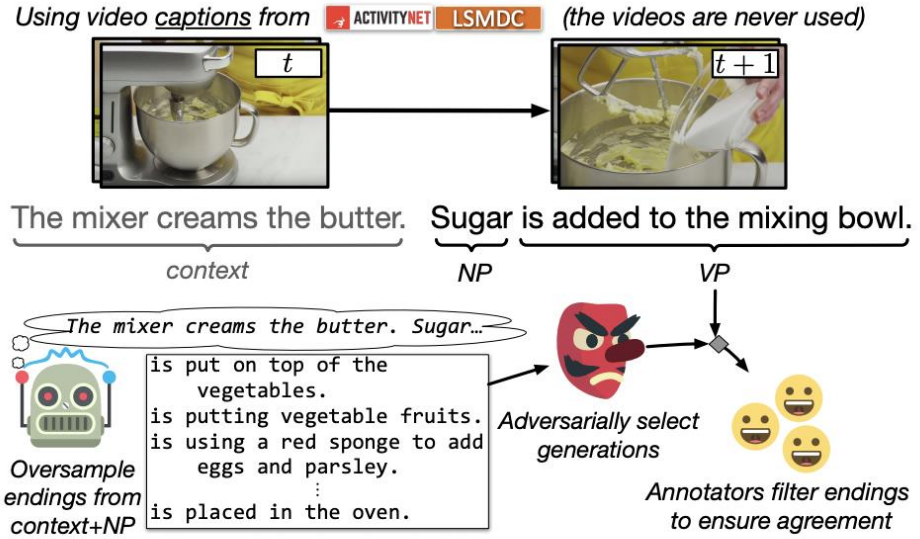


Figure 1: Overview of the data collection process. For a pair of sequential video captions, the second caption is split into noun and verb phrases. A language model generates many negative endings, of which a difficult subset are human-annotated.

Zellers, R., Bisk, Y., Schwartz, R., & Choi, Y. (2018). Swag: A large-scale adversarial dataset for grounded commonsense inference. arXiv:1808.0532

WinoGrande

- robust commonsense capabilities or rely on spurious biases (with ✗ in the example below)
- improve both the scale and the hardness of the WSC

		Twin sentences	Options (**answer**)
✓ (1)	a	The trophy doesn't fit into the brown suitcase because **it**'s too *large*.	**trophy** / suitcase
	b	The trophy doesn't fit into the brown suitcase because **it**'s too *small*.	trophy / **suitcase**
✓ (2)	a	Ann asked Mary what time the library closes, *because* **she** had forgotten.	**Ann** / Mary
	b	Ann asked Mary what time the library closes, *but* **she** had forgotten.	Ann / **Mary**
✗ (3)	a	The tree fell down and crashed through the roof of my house. Now, I have to get **it** *removed*.	**tree** / roof
	b	The tree fell down and crashed through the roof of my house. Now, I have to get **it** *repaired*.	tree / **roof**
✗ (4)	a	The lions ate the zebras because **they** are *predators*.	**lions** / zebras
	b	The lions ate the zebras because **they** are *meaty*.	lions / **zebras**

Sakaguchi, K., Bras, R. L., Bhagavatula, C., & Choi, Y. (2019). WINOGRANDE: An adversarial winograd schema challenge at scale.

WinoGrande AFLITE

- adopt a dense representation of instances using precomputed neural network embeddings
- an ensemble of linear classifiers (logistic regressions) trained on random subsets of the

dataset-specific bias detected by AFLITE (marked with ✗)

	Twin sentences	Options (**answer**)
✗	The monkey loved to play with the balls but ignored the blocks because he found **them** *exciting*.	**balls** / blocks
	The monkey loved to play with the balls but ignored the blocks because he found **them** *dull*.	balls / **blocks**
✗	William could only climb begginner walls while Jason climbed advanced ones because **he** was very *weak*.	**William** / Jason
	William could only climb begginner walls while Jason climbed advanced ones because **he** was very *strong*.	William / **Jason**
✓	Robert woke up at 9:00am while Samuel woke up at 6:00am, so **he** had *less* time to get ready for school.	**Robert** / Samuel
	Robert woke up at 9:00am while Samuel woke up at 6:00am, so **he** had *more* time to get ready for school.	Robert / **Samuel**
✓	The child was screaming after the baby bottle and toy fell. Since the child was *hungry*, **it** stopped his crying.	**baby bottle** / toy
	The child was screaming after the baby bottle and toy fell. Since the child was *full*, **it** stopped his crying.	baby bottle / **toy**

Sakaguchi, K., Bras, R. L., Bhagavatula, C., & Choi, Y. (2019). WINOGRANDE: An adversarial winograd schema challenge at scale.

TextFooler

1. Word Importance Ranking
2. Word Transformer (replacement)
   - have similar semantic meaning with the original
   - fit within the surrounding context
   - force the target model to make wrong predictions

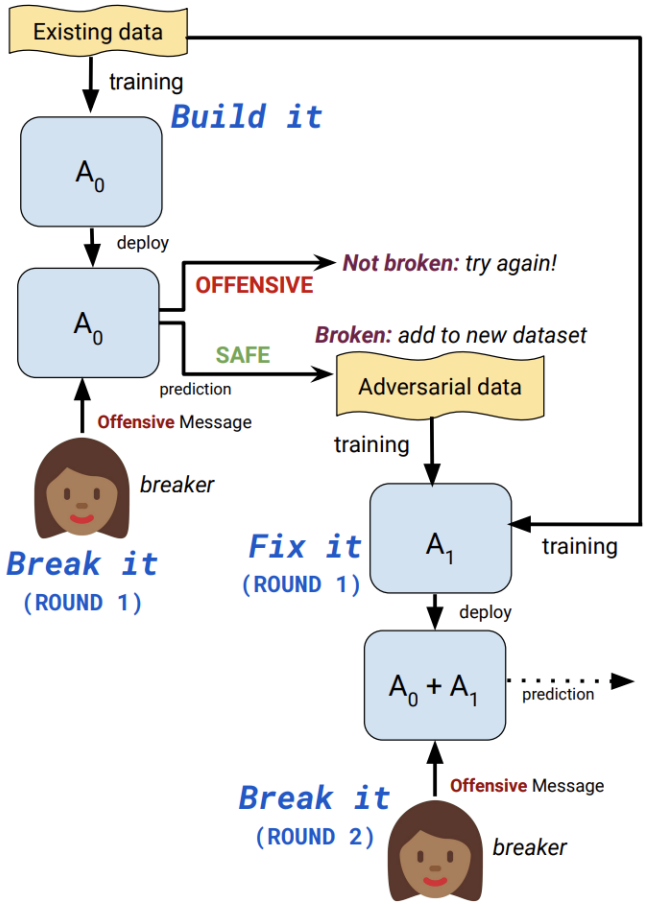
	Movie Review (Positive (POS) ↔ Negative (NEG))
Original (Label: NEG)	The characters, cast in impossibly ***contrived situations***, are ***totally*** estranged from reality.
Attack (Label: POS)	The characters, cast in impossibly ***engineered circumstances***, are ***fully*** estranged from reality.
Original (Label: POS)	It cuts to the ***knot*** of what it actually means to face your ***scares***, and to ride the ***overwhelming*** **metaphorical wave** that life wherever it takes you.
Attack (Label: NEG)	It cuts to the ***core*** of what it actually means to face your ***fears***, and to ride the ***big*** **metaphorical wave** that life wherever it takes you.
	SNLI (Entailment (ENT), Neutral (NEU), Contradiction (CON))
Premise	Two small boys in blue soccer uniforms use a wooden set of steps to wash their hands.
Original (Label: CON)	The boys are in band ***uniforms***.
Adversary (Label: ENT)	The boys are in band ***garment***.
Premise	A child with wet hair is holding a butterfly decorated beach ball.
Original (Label: NEU)	The ***child*** is at the ***beach***.
Adversary (Label: ENT)	The ***youngster*** is at the ***shore***.

Table 6: Examples of original and adversarial sentences from MR (WordLSTM) and SNLI (BERT) datasets.

Di Jin (2019). Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment. arXiv:1907.11932

Build it
Break it
Fix it

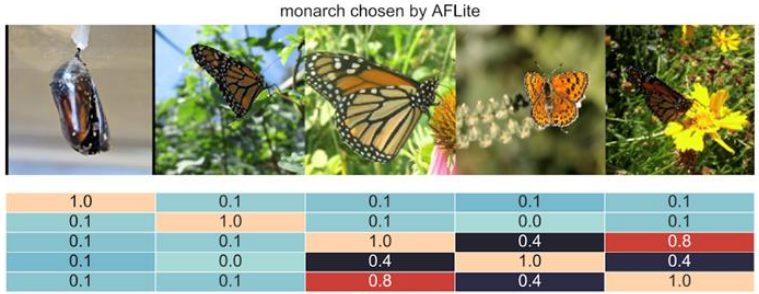
- a training scheme for a model to become robust
- iterative build it, break it, fix it strategy with humans and models in the loop

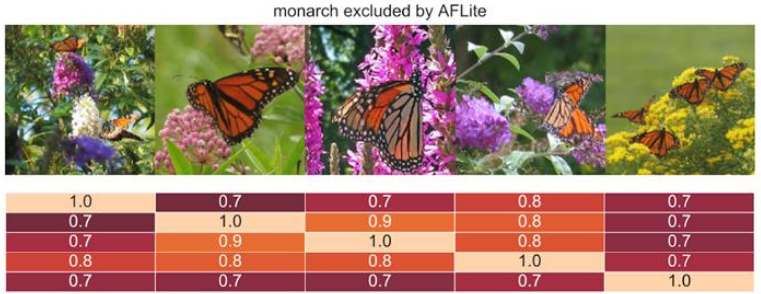


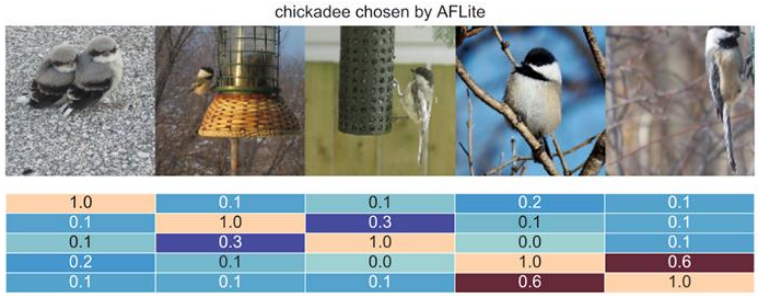
Dinan, E., Humeau, S., Chintagunta, B., & Weston, J. (2019). Build it break it fix it for dialogue safety: Robustness from adversarial human attack. arXiv:1908.06083.

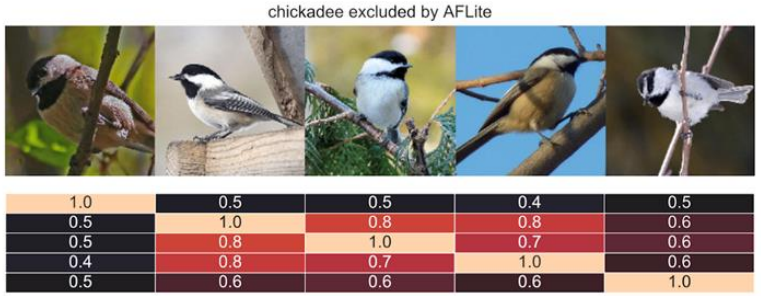
AFLite Investigation

- provide a theoretical understanding
- proves models trained on the filtered datasets yield better generalization









Bras, R. L., Swayamdipta, S., Bhagavatula, C., Zellers, R., Peters, M. E., Sabharwal, A., & Choi, Y. (2020). Adversarial Filters of Dataset Biases. arXiv:2002.04108

But wait! There’s one more thing

Accuracy isn’t everything

Accuracy is not the direct measure for robustness.

Consistency is!

Definition of consistency:
Question A and A’ are a dual test pair
A consistent case would be:
Model get both A and A’ right or wrong

A: He drinks apple.

A’: It is he who drinks apple.

Consistency and accuracy are not the same.

Model	Full WSC Acc.	Unswitched Acc.	Switched Acc.	Consistency
Single LM	54.58%	54.96%	54.20%	56.49%
Ensemble 10 LMs	61.54%	58.78%	49.62%	43.51%
Ensemble 14 LMs	63.74%	63.36%	53.43%	44.27%
GPT-2 117M Full	55.68%	54.20%	54.20%	26.72%
GPT-2 117M Partial	61.54%	59.54%	52.67%	48.85%
GPT-2 774M Full	64.47%	62.60%	54.96%	45.04%
GPT-2 774M Partial	69.23%	67.94%	61.83%	63.35%
Knowledge Hunter	57.14%[4]	58.78%[4]	58.78%[4]	90.07%[4]

Table 2: Evaluation of state-of-the-art methods on WSC using the proposed switchability metrics. The last three columns give numbers on the switchable subset only.

Trichelair, P., Emami, A., Trischler, A., Suleman, K., & Cheung, J. C. K. (2018). How Reasonable are Common-Sense Reasoning Tasks: A Case-Study on the Winograd Schema Challenge and SWAG.

Consistency and accuracy are not the same.

(3) **Original:** *Emma* did not pass the ball to *Janie* although she saw that she was open.

(4) **Switched:** *Janie* did not pass the ball to *Emma* although she saw that she was open.

Trichelair, P., Emami, A., Trischler, A., Suleman, K., & Cheung, J. C. K. (2018). How Reasonable are Common-Sense Reasoning Tasks: A Case-Study on the Winograd Schema Challenge and SWAG.

Our next step towards final project

Measure the consistency of BERT & GPT-2 using our own dataset