Back to Square One: Artifact Detection, Training and Commonsense Disentanglement in the Winograd Schema

Yanai Elazar, Hongming Zhang,
Yoav Goldberg, Dan Roth

EMNLP 2021

Commonsense Reasoning

Wikipedia Definition for:

Commonsense reasoning is one of the branches of artificial intelligence (AI) that is concerned with simulating the human ability to make presumptions about the type and essence of ordinary situations they encounter every day.

Commonsense Reasoning

That is

Commonsense Reasoning

That is

- Someone passes through a door → they are smaller than it

Commonsense Reasoning

That is

- Someone passes through a door → they are smaller than it
- It's 11:00 → Need to order food

Commonsense Reasoning

That is

- Someone passes through a door → they are smaller than it
- It’s 11:00 → Need to order food
- I’m giving a talk today → I should probably start preparing the slides

Commonsense Reasoning

Meanwhile, in NLP

GPT-3

Generative Pre-trained Transformer 3 is an autoregressive language model that uses deep learning to produce human-like text. It is the third-generation language prediction model in the GPT-n series created by OpenAI, a San Francisco-based artificial intelligence research laboratory. Wikipedia

Original author: OpenAI

Initial release: June 11, 2020 (beta)

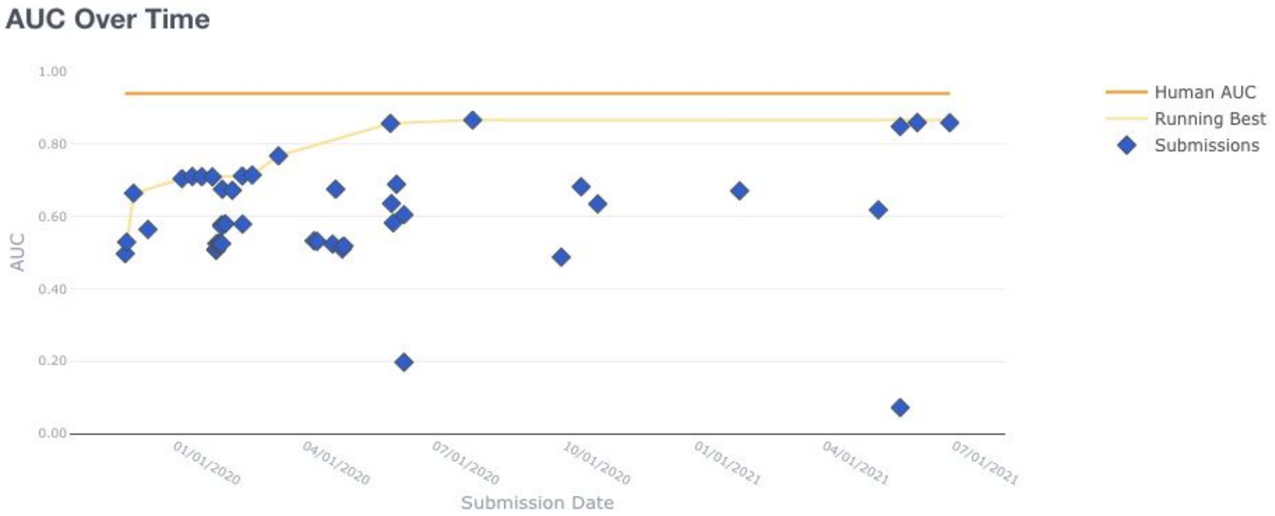
License: Code unavailable, only accessible by a paywalled API

Feedback

Meanwhile, in NLP

WinoGrande: An Adversarial Winograd Schema Challenge at Scale

Keisuke Sakaguchi*, Ronan Le Bras*, Chandra Bhagavatula*, Yejin Choi*†
*Allen Institute for Artificial Intelligence †University of Washington
{keisukes, ronanlb, chandrab, yejinc}@allenai.org



Meanwhile, in NLP

WinoGrande: An Adversarial Winograd Schema Challenge at Scale

Keisuke Sakaguchi*, Ronan Le Bras*, Chandra Bhagavatula*, Yejin Choi*†
*Allen Institute for Artificial Intelligence †University of Washington
{keisukes, ronanlb, chandrab, yejinc}@allenai.org

<u>Assumption</u>:
Main reason for commonsense reasoning improvement is due to better LMs

Commonsense Reasoning Through the Winograd Schema

The Winograd Schema

- Introduced in 2011 as an alternative to the Turing Test by Hector J. Levesque
- The purpose is to test for common sense
- *"... Moreover, the test is arranged in such a way that having full access to a large corpus of English text might not help much ..."*

The Winograd Schema

Every question involves:

Joan made sure to thank Susan for all the help she had given.

1. Two entities are mentioned in each sentence, and they can be two males, two females, two inanimate objects, or two groups of people or objects;

The Winograd Schema

Every question involves:

Joan *made sure to thank* ***Susan*** *for all the help she had given.*

1. Two entities are mentioned in each sentence, and they can be two males, two females, two inanimate objects, or two groups of people or objects;

The Winograd Schema

Every question involves:

Joan made sure to thank Susan for all the help she had given.

1. Two entities are mentioned in each sentence, and they can be two males, two females, two inanimate objects, or two groups of people or objects;
2. A pronoun is used in the example to refer to one of the entities

The Winograd Schema

Every question involves:

Joan made sure to thank Susan for all the help ***she*** *had given.*

1. Two entities are mentioned in each sentence, and they can be two males, two females, two inanimate objects, or two groups of people or objects;
2. A pronoun is used in the example to refer to one of the entities

The Winograd Schema

Every question involves:

Joan made sure to thank Susan for all the help she had given.

1. Two entities are mentioned in each sentence, and they can be two males, two females, two inanimate objects, or two groups of people or objects;
2. A pronoun is used in the example to refer to one of the entities
3. The task is to determine which of the two entities is referred to by the pronoun (coreference)

The Winograd Schema

Every question involves:

***Joan** made sure to thank **Susan** for all the help **she** had given.*

1. Two entities are mentioned in each sentence, and they can be two males, two females, two inanimate objects, or two groups of people or objects;
2. A pronoun is used in the example to refer to one of the entities
3. The task is to determine which of the two entities is referred to by the pronoun (coreference)

The Winograd Schema

Every question involves:

Joan made sure to thank Susan for all the help she had given.

1. Two entities are mentioned in each sentence, and they can be two males, two females, two inanimate objects, or two groups of people or objects;
2. A pronoun is used in the example to refer to one of the entities
3. The task is to determine which of the two entities is referred to by the pronoun (coreference)
4. Each sentence contains a **special word** which, when replaced, the answer changes.

The Winograd Schema

Every question involves:

Joan made sure to thank Susan for all the help she had <u>given</u>.

1. Two entities are mentioned in each sentence, and they can be two males, two females, two inanimate objects, or two groups of people or objects;
2. A pronoun is used in the example to refer to one of the entities
3. The task is to determine which of the two entities is referred to by the pronoun (coreference)
4. Each sentence contains a **special word** which, when replaced, the answer changes.

The Winograd Schema

- ***Joan*** *made sure to thank* ***Susan*** *for all the help* ***she*** *had* *given*.

The Winograd Schema

- ***Joan*** *made sure to thank* ***Susan*** *for all the help* ***she*** *had* *<u>given</u>*.
- ***Joan*** *made sure to thank* ***Susan*** *for all the help* ***she*** *had* *<u>received</u>*.

The Winograd Schema

- ***Joan*** *made sure to thank* ***Susan*** *for all the help* ***she*** *had <u>given</u>.*
- ***Joan*** *made sure to thank* ***Susan*** *for all the help* ***she*** *had <u>received</u>.*

The Winograd Schema

- ***Joan*** *made sure to thank* ***Susan*** *for all the help* ***she*** *had <u>given</u>.*
- ***Joan*** *made sure to thank* ***Susan*** *for all the help* ***she*** *had <u>received</u>.*

- *The* ***trophy*** *doesn't fit in the brown* ***suitcase*** *because* ***it*** *was too <u>large</u>.*

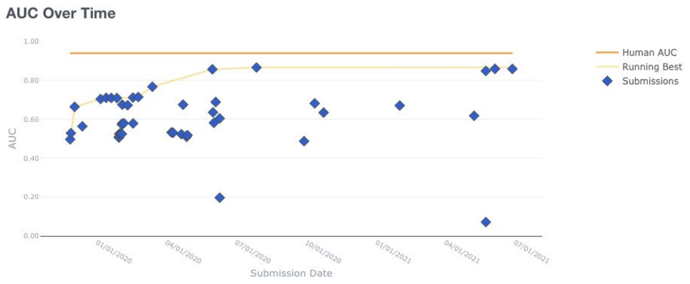
The Winograd Schema

- ***Joan*** *made sure to thank* ***Susan*** *for all the help* ***she*** *had <u>given</u>.*
- ***Joan*** *made sure to thank* ***Susan*** *for all the help* ***she*** *had <u>received</u>.*

- *The* ***trophy*** *doesn't fit in the brown* ***suitcase*** *because* ***it*** *was too <u>large</u>.*
- *The* ***trophy*** *doesn't fit in the brown* ***suitcase*** *because* ***it*** *was too <u>small</u>.*

The Winograd Schema

- Initial dataset of 273 examples *--Levesque et al., 2012*
    - Written by experts
- 2 years ago: Winogrande with 44K examples *--Sakaguchi et al., 2019*
    - Written by crowdworkers
- Today:



3 Reasons Why...
Winograd Schema Results are Inflated

1. Artifacts
2. Evaluation
3. Limited Generalization

Artifacts in the Data

The Winograd Schema - Artifacts?

- Signals that can help solving the problem without the expected type of inference
    - *The **racecar** zoomed by the **school bus** because* ***it*** *was going so* *<u>fast</u>*.
- *We design two methods to discover such artifacts*

Artifacts Discovery: No-Candidates

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

Artifacts Discovery: No-Candidates

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

Artifacts Discovery: Part-Sentences

- *The **trophy** doesn't fit into the brown **suitcase** because **it** is too large.*

Artifacts Discovery: Part-Sentences

- ~~*The **trophy** doesn't fit into the brown **suitcase***~~ *because **it** is too large.*
- *because **it** is too large.*

Reminiscent of *Trichelair et al. 2019*

Artifacts Discovery: Results

Setup:

- Training a model on Winogrande, a large (44K) crowdsourced dataset for the winograd schema.
    - Each sentence is replaced with each entity, then a score is calculated for each alternative
        - The **trophy** doesn't fit into the brown **suitcase** because the **trophy** is too <u>large</u>.
        - The **trophy** doesn't fit into the brown **suitcase** because the **suitcase** is too <u>large</u>.

Artifacts Discovery: Results

Setup:

- Training a model on Winogrande, a large (44K) crowdsourced dataset for the winograd schema.
    - Each sentence is replaced with each entity, then a score is calculated for each alternative
        - ~~The **trophy** doesn't fit into the brown **suitcase**~~ because the **trophy** is too <u>large</u>.
        - ~~The **trophy** doesn't fit into the brown **suitcase**~~ because the **suitcase** is too <u>large</u>.

Artifacts Discovery: Results

Setup:

- Training a model on Winogrande, a large (44K) crowdsourced dataset for the winograd schema.
  - Each sentence is replaced with each entity, then a score is calculated for each alternative
    - The **trophy** doesn't fit into the brown **suitcase** because the **trophy** is too <u>large</u>.
    - The **trophy** doesn't fit into the brown **suitcase** because the **suitcase** is too <u>large</u>.
- Test the trained model on the different setups

Artifacts Discovery: Results

Dataset	Setup	Single
-	random	50.0
WSC	original	89.71
	no-cands	60.72
	part-sent	64.88
WSC-na	original	89.45
	no-cands	58.06
	part-sent	59.90
Winogrande	original	71.49
	no-cands	53.07
	part-sent	53.11

Artifacts Discovery: Results

Dataset	Setup	Single
-	random	50.0
WSC	original	89.71
	no-cands	60.72
	part-sent	64.88
WSC-na	original	89.45
	no-cands	58.06
	part-sent	59.90
Winogrande	original	71.49
	no-cands	53.07
	part-sent	53.11

Artifacts Discovery: Results

Dataset	Setup	Single
-	random	50.0
WSC	original	89.71
	no-cands	60.72
	part-sent	64.88
WSC-na	original	89.45
	no-cands	58.06
	part-sent	59.90
Winogrande	original	71.49
	no-cands	53.07
	part-sent	53.11

Artifacts Discovery: Results

Dataset	Setup	Single
-	random	50.0
WSC	original	89.71
	no-cands	60.72
	part-sent	64.88
WSC-na	original	89.45
	no-cands	58.06
	part-sent	59.90
Winogrande	original	71.49
	no-cands	53.07
	part-sent	53.11

Artifacts Discovery: Results

Dataset	Setup	Single
-	random	50.0
WSC	original	89.71
	no-cands	60.72
	part-sent	64.88
WSC-na	original	89.45
	no-cands	58.06
	part-sent	59.90
Winogrande	original	71.49
	no-cands	53.07
	part-sent	53.11

>> *random!*

Artifacts Discovery: Results

Dataset	Setup	Single
-	random	50.0
WSC	original	89.71
	no-cands	60.72
	part-sent	64.88
WSC-na	original	89.45
	no-cands	58.06
	part-sent	59.90
Winogrande	original	71.49
	no-cands	53.07
	part-sent	53.11

~ *random*

Artifacts Discovery: Results

Dataset	Setup	Single
	random	50.0
WSC	original	89.71
	no-cands	60.72
	part-sent	64.88
WSC-na	original	[illegible]
	no-cands	58.06
	part-sent	59.90
Winogrande	original	71.49
	no-cands	53.07
	part-sent	53.11

Human experts may leak artifacts into the data

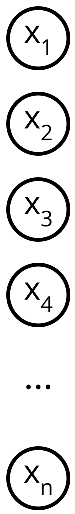
WSC suffer from some artifacts
Winogrande - less so

Evaluation

Part II

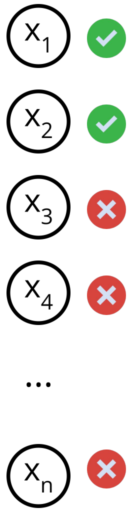
Evaluation up to date

- We get a set of inputs, and report accuracy



Evaluation up to date

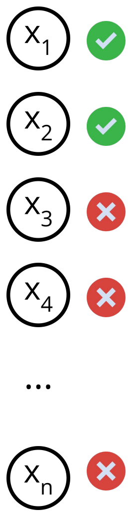
- We get a set of inputs, and report accuracy



= 7 / 10 = 70%

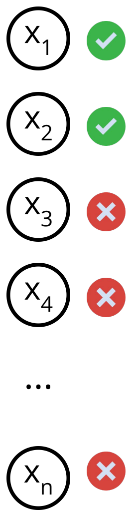
Evaluation up to date

- We get a set of inputs, and report accuracy
- and this is fine, when the data is sampled i.i.d



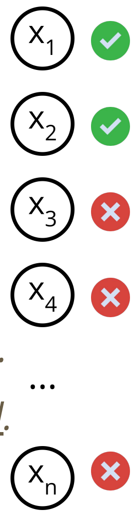
Evaluation up to date

- We get a set of inputs, and report accuracy
- and this is fine, when the data is sampled i.i.d
- But this is not the case in the winograd schema!



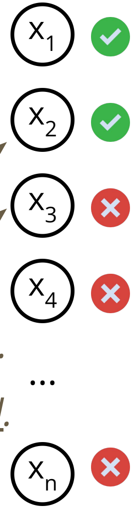
Evaluation up to date

- We get a set of inputs, and report accuracy
- and this is fine, when the data is sampled i.i.d
- But this is not the case in the winograd schema!
- Recall the pairs:
    - *The **trophy** doesn't fit into the brown **suitcase** because it is too large.*
    - *The **trophy** doesn't fit into the brown **suitcase** because it is too small.*



Evaluation up to date

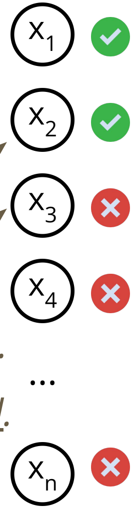
- We get a set of inputs, and report accuracy
- and this is fine, when the data is sampled i.i.d
- But this is not the case in the winograd schema!
- Recall the pairs:
    - *The **trophy** doesn't fit into the brown **suitcase** because it is too large.*
    - *The **trophy** doesn't fit into the brown **suitcase** because it is too small.*



Evaluation up to date

- We get a set of inputs, and report accuracy
- and this is fine, when the data is sampled i.i.d
- But this is not the case in the winograd schema!
- Recall the pairs:
    - *The* ***trophy*** *doesn't fit into the brown* ***suitcase*** *because it is too large.*
    - *The* ***trophy*** *doesn't fit into the brown* ***suitcase*** *because it is too small.*
- If a model got only one item of a pair right,

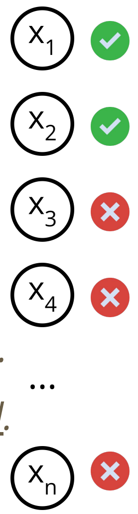
  did it really understand the question?



Evaluation up to date

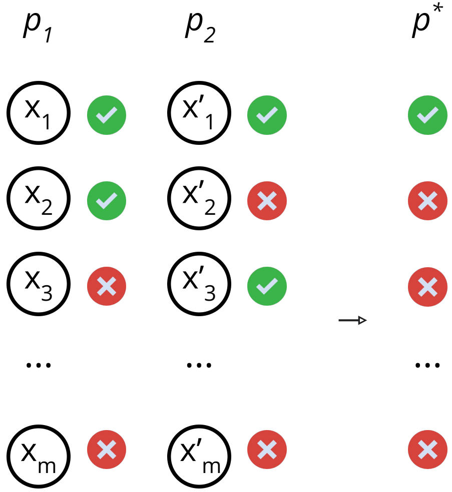
- We get a set of inputs, and report accuracy
- and this is fine, when the data is sampled i.i.d
- But this is not the case in the winograd schema!
- Recall the pairs:
    - *The **trophy** doesn't fit into the brown **suitcase** because it is too large.*
    - *The **trophy** doesn't fit into the brown **suitcase** because it is too small.*
- If a model got only one item of a pair right,

  did it really understand the question?
    - **No!** This results from **randomness**, or **artifacts** in the data



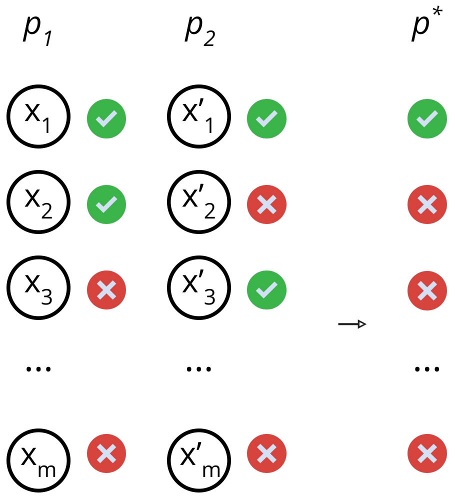
Paired Evaluation

- Instead, let's assign a point to a pair, only if a model gets both right



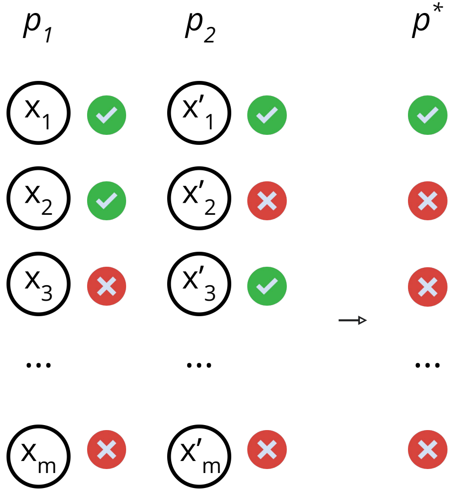
Paired Evaluation

- Instead, let's assign a point to a pair, only if a model gets both right
- This way, the risk of giving away points is reduced…



Paired Evaluation

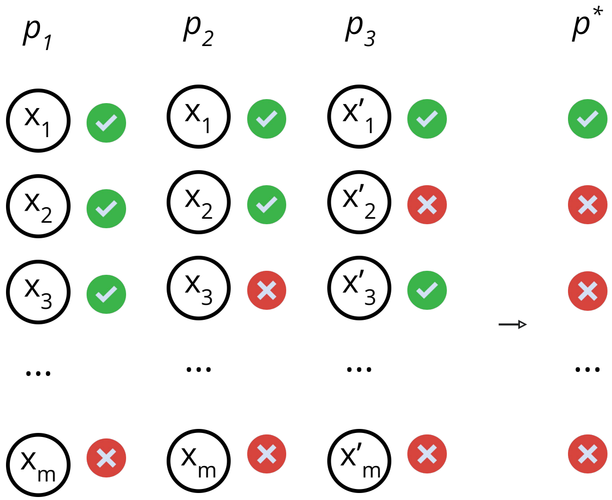
- Instead, let's assign a point to a pair, only if a model gets both right
- This way, the risk of giving away points is reduced…
- and this evaluation becomes more **robust** and **meaningful**



Group Evaluation

- We also generalize this evaluation to groups and an arbitrary function

$$groupScore(x_i) = \min_j f(x_{i_j})$$



Group Evaluation

Dataset	Setup	Single	Group
-	random	50.0	25.0
WSC	original	89.71	79.41
	no-cands	60.72	40.35
	part-sent	64.88	33.88
WSC-na	original	89.45	79.09
	no-cands	58.06	34.41
	part-sent	59.90	25.00
Winogrande	original	71.49	58.45
	no-cands	53.07	31.05
	part-sent	53.11	22.34

Group Evaluation

Dataset	Setup	Single	Group
-	random	50.0	25.0
WSC	original	89.71	79.41
	no-cands	60.72	40.35
	part-sent	64.88	33.88
WSC-na	original	89.45	79.09
	no-cands	58.06	34.41
	part-sent	59.90	25.00
Winogrande	original	71.49	58.45
	no-cands	53.07	31.05
	part-sent	53.11	22.34

Group Evaluation

Dataset	Setup	Single	Group
-	random	50.0	25.0
WSC	original	89.71	79.41
	no-cands	60.72	40.35
	part-sent	64.88	33.88
WSC-na	original	89.45	79.09
	no-cands	58.06	34.41
	part-sent	59.90	25.00
Winogrande	original	71.49	58.45
	no-cands	53.07	31.05
	part-sent	53.11	22.34

Group Evaluation

Dataset	Setup	Single	Group
-	random	50.0	25.0
WSC	original	89.71	79.41
	no-cands	60.72	40.35
	part-sent	64.88	33.88
WSC-na	original	89.45	79.09
	no-cands	58.06	34.41
	part-sent	59.90	25.00
Winogrande	original	71.49	58.45
	no-cands	53.07	31.05
	part-sent	53.11	22.34

Knowledge and Format Disentanglement

Part III

Commonsense Reasoning Training

LMs trained on Winogrande are getting close to human agreement on the Winograd schema

Commonsense Reasoning Training

LMs trained on Winogrande are getting close to human agreement on the Winograd schema

But WAIT!

Commonsense Reasoning Training

LMs trained on Winogrande are getting close to human agreement on the Winograd schema

But WAIT!

Do we even want to train on such dataset?

Commonsense Reasoning Training

- Limited generalization
  - Learning about the strength of steel would teach a model about the strength of wood? And about the strength of styrofoam?
- The commonsense space is huge, it is not reasonable to learn it from a limited dataset

Let’s measure progress in a zero-shot setting

Let MLM Do MLM

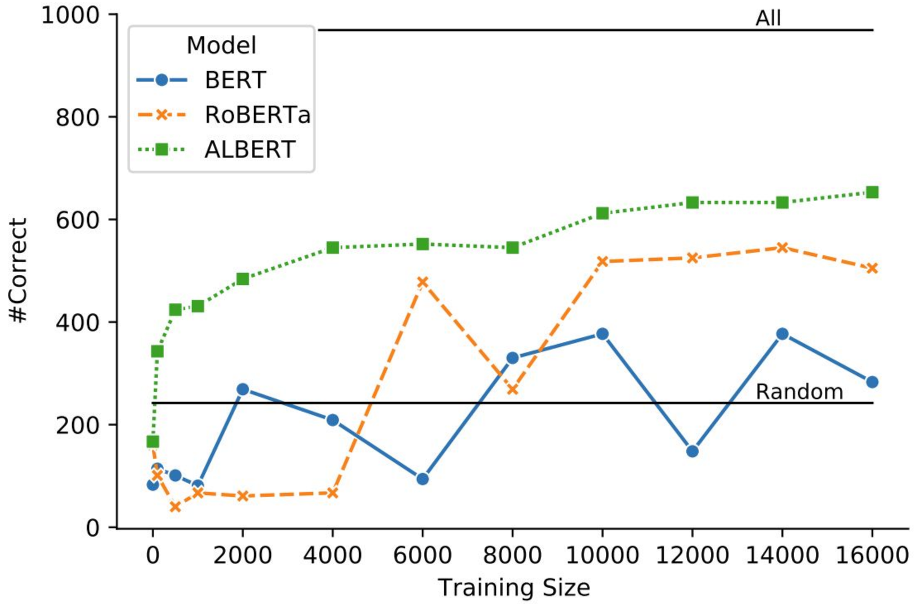
- Previous methods for measuring zero-shot performance using LMs are flawed
- We propose a new method which allows us to properly measure it (more details in the paper)

Let MLM Do MLM - Zero Shot Evaluation

What does it mean?

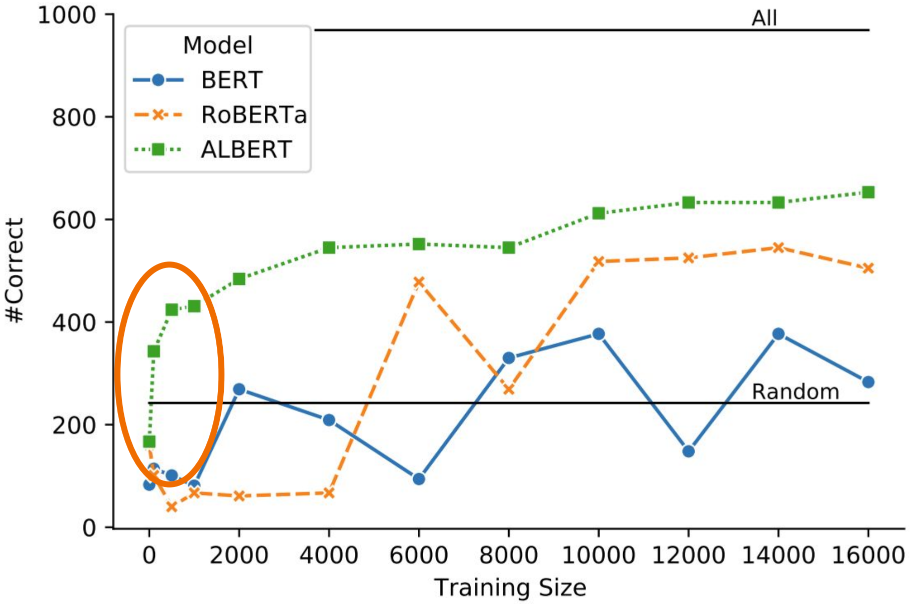
Model	WinoGrande Single	WinoGrande Group
random	50.00	25.00
BERT-base	53.12	11.11
BERT-large	55.56	12.50
RoBERTa-base	56.25	14.58
RoBERTa-large	54.86	12.50
ALBERT-base	52.78	7.64
ALBERT-xxlarge	58.68	20.83

Pre-Trained Models: From Hero to Zero



Pre-Trained Models: From Hero to Zero

- Finetuning contribute to the #correct predictions **slightly**
- This suggests that the supervision for WS commonsense reasoning is **merely** beneficial and it is hard to generalize



What's Next?

- Decoupling commonsense knowledge

What's Next?

- Decoupling commonsense knowledge from reasoning

What's Next?

- Decoupling commonsense knowledge from reasoning
- Can we teach the reasoning? (similar to *Clark et al. 2020*)

What's Next?

- Decoupling commonsense knowledge from reasoning
- Can we teach the reasoning? (similar to *Clark et al. 2020*)
- Rigorous definitions for commonsense generalizations

Summary

- Automatic control baselines measuring artifacts in WS data
- *Group-Scoring*: a more robust evaluation for minimal-distance groups
- Zero-shot evaluation for WS
- Results indicate that the **progress does not come from better LMs, but from data, which should be used for evaluation, not training**

Thanks!

Questions?