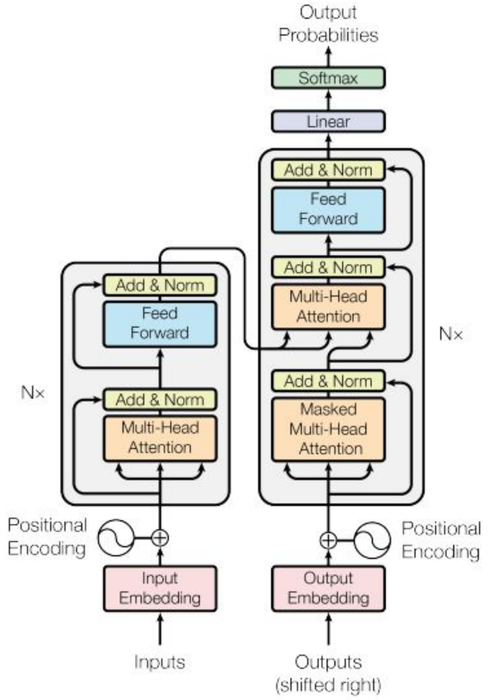
Complex Commonsense Reasoning

Why we are not there yet

Future of AI: Yanai Elazar

State of The Art NLP

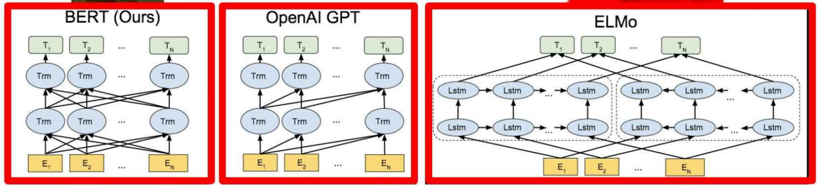
A Deep Neural Network (2017), called:
Transformers

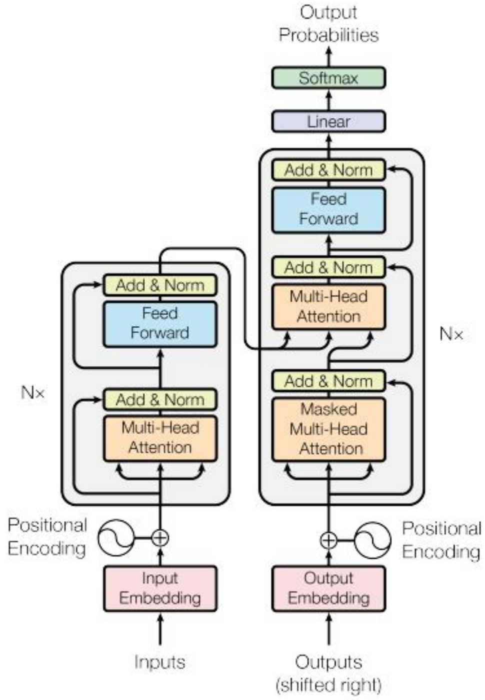


State of the Art NLP

Since 2018, using the **Transformers**
to train Big *Language Models*
to predict words in context

BERT

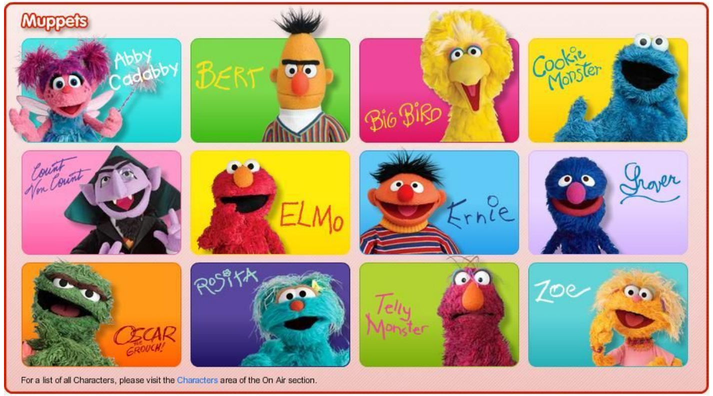




State of The Art NLP

NLP
since 2018

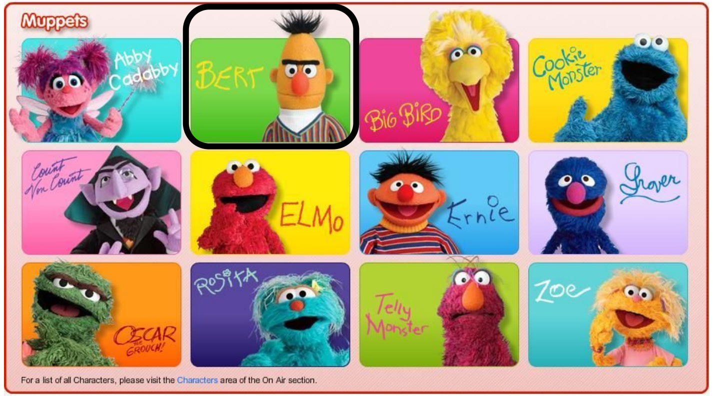
Step 1: Pick your favorite muppet



State of The Art NLP

NLP
since 2018

Step 1: Pick your favorite muppet

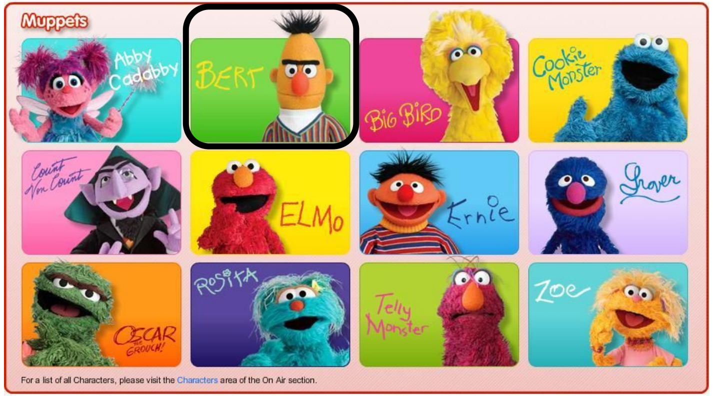


State of The Art NLP

NLP
since 2018

Step 1: Pick your favorite muppet

Step 2: Bring your own data + Train



Datasets

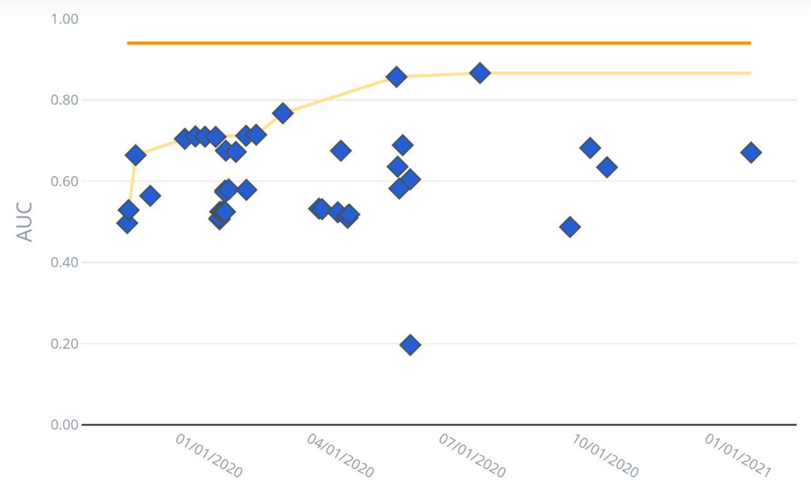
State of The Art NLP

NLP
since 2018

Step 1: Pick your favorite muppet

Step 2: Bring your own data + Train

Step 3: Rock the leaderboard



Datasets

State of the Art NLP

NLP
since 2018

GLUE

Rank	Name	Model	URL	Score
1	ERNIE Team - Baidu	ERNIE		90.9
2	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.8
3	HFL iFLYTEK	MacALBERT + DKM		90.7
4	Alibaba DAMO NLP	StructBERT + TAPT		90.6
5	PING-AN Omni-Sinitic	ALBERT + DAAF + NAS		90.6
6	T5 Team - Google	T5		90.3
7	Microsoft D365 AI & MSR AI & GATECH	MT-DNN-SMART		89.9
8	Huawei Noah's Ark Lab	NEZHA-Large		89.8
9	Zihang Dai	Funnel-Transformer (Ensemble B10-10-10H1024)		89.7
10	ELECTRA Team	ELECTRA-Large + Standard Tricks		89.4
11	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.4
12	Junjie Yang	HIRE-RoBERTa		88.3
13	Facebook AI	RoBERTa		88.1
14	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6
15	GLUE Human Baselines	GLUE Human Baselines		87.1
16	Adrian de Wynter	Bort (Alexa AI)		83.6
17	Lab LV	ConvBERT base		83.2
18	Stanford Hazy Research	Snorkel MeTaL		83.2

SuperGLUE

Rank	Name	Model	URL	Score
1	Zirui Wang	T5 + Meena, Single Model (Meena Team - Google Brain)		90.4
2	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.3
3	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8
4	T5 Team - Google	T5		89.3
5	Huawei Noah's Ark Lab	NEZHA-Plus		86.7
6	Alibaba PAI&ICBU	PAI Albert		86.1
7	Infosys : DAWN : AI Research	RoBERTa-iCETS		86.0
8	Tencent Jarvis Lab	RoBERTa (ensemble)		85.9
9	Zhuiyi Technology	RoBERTa-mtl-adv		85.7
10	Facebook AI	RoBERTa		84.6
11	Anuar Sharafudinov	AILabs Team Transformers		77.5
12	Timo Schick	iPET (ALBERT) - Few-Shot (32 Examples)		75.4
13	Adrian de Wynter	Bort (Alexa AI)		74.1
14	IBM Research AI	BERT-mtl		73.5
15	Ben Mann	GPT-3 few-shot - OpenAI		71.8
16	SuperGLUE Baselines	BERT++		71.5
		BERT		69.0

Can We Go Home??

(No)

Case Study

Commonsense Reasoning

Supervised training is not always the answer

The Winograd Schema

- Introduced in 2011 as an alternative to the Turing Test by Hector J. Levesque

- "... 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

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

The Winograd Schema

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

The Winograd Schema

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

The Winograd Schema

- The **trophy** doesn’t fit in the brown **suitcase** because it was too large.

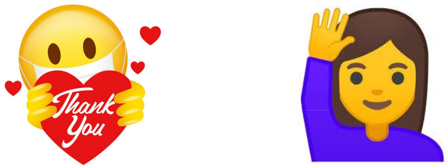
The Winograd Schema

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

The Winograd Schema

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

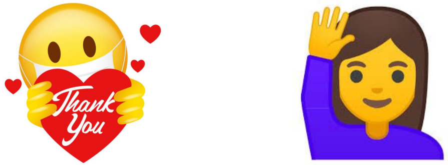
- **Joan** made sure to thank **Susan** for all the help she had given.



The Winograd Schema

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

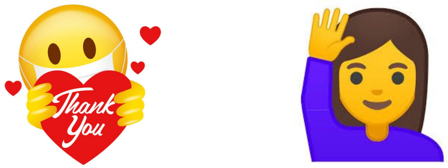
- **Joan** made sure to thank **Susan** for all the help she had given.



The Winograd Schema

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

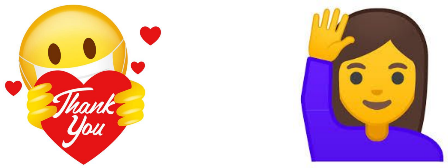
- **Joan** made sure to thank **Susan** for all the help she had given.



The Winograd Schema

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

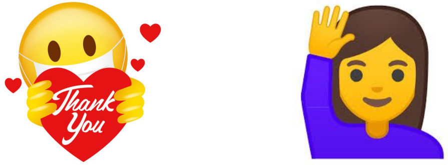
- **Joan** made sure to thank **Susan** for all the help she had given.



The Winograd Schema

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

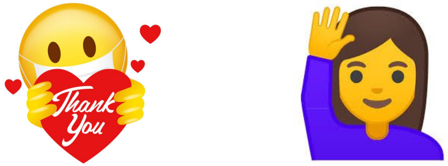
- **Joan** made sure to thank **Susan** for all the help she had given.



The Winograd Schema

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

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



Why is it hard?

The Winograd Schema

Every question in the schema involves 4 key points:

1. Two entities are mentioned in each sentence

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

The Winograd Schema

Every question in the schema involves 4 key points:

1. Two entities are mentioned in each sentence

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

The Winograd Schema

Every question in the schema involves 4 key points:

1. Two entities are mentioned in each sentence
2. A pronoun is used in the example to refer to one of the entities

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

The Winograd Schema

Every question in the schema involves 4 key points:

1. Two entities are mentioned in each sentence
2. A pronoun is used in the example to refer to one of the entities

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

The Winograd Schema

Every question in the schema involves 4 key points:

1. Two entities are mentioned in each sentence
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)

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

The Winograd Schema

Every question in the schema involves 4 key points:

1. Two entities are mentioned in each sentence
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)

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

The Winograd Schema

Every question in the schema involves 4 key points:

1. Two entities are mentioned in each sentence
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.

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

The Winograd Schema

Every question in the schema involves 4 key points:

1. Two entities are mentioned in each sentence
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.

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

The Winograd Schema

Every question in the schema involves 4 key points:

1. Two entities are mentioned in each sentence
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.

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

The Winograd Schema

Every question in the schema involves 4 key points:

1. Two entities are mentioned in each sentence
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.

received

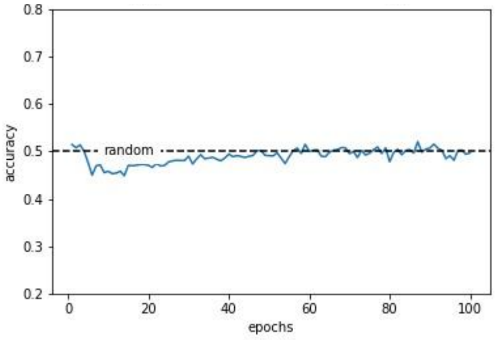
Joan made sure to thank Susan for all the help she had ~~given~~.

The Winograd Schema

- Was considered a hard task for years

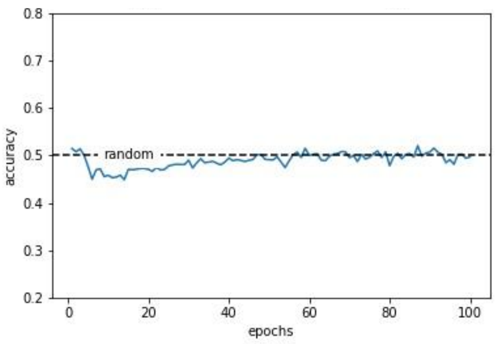
The Winograd Schema

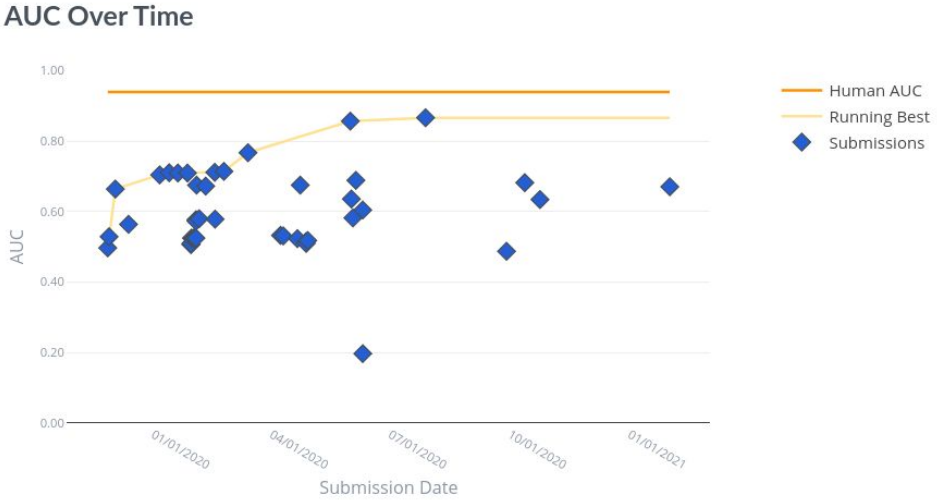
- Was considered a hard task for years
- Until recently, models' performance oscillated near random



The Winograd Schema

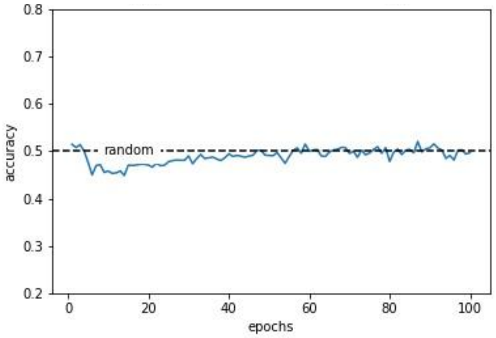
- Was considered a hard task for years
- Until recently, models' performance oscillated near random
- But now things looks different

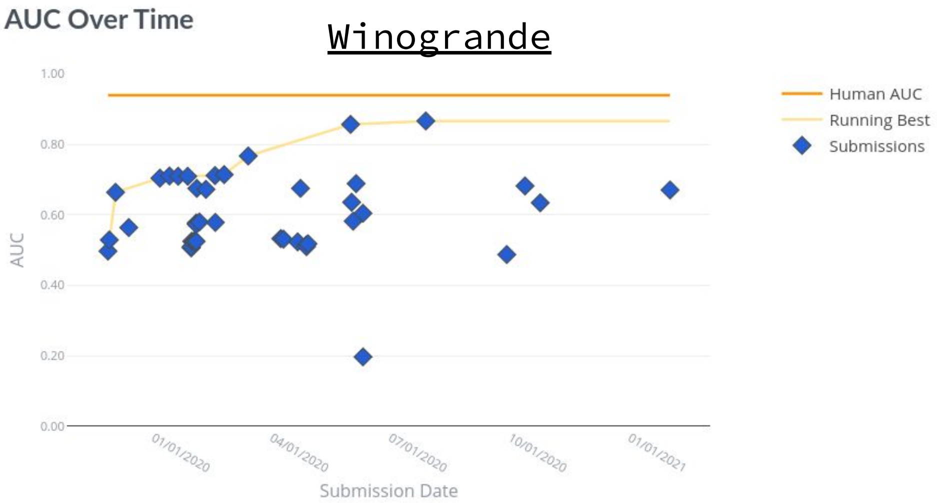




The Winograd Schema

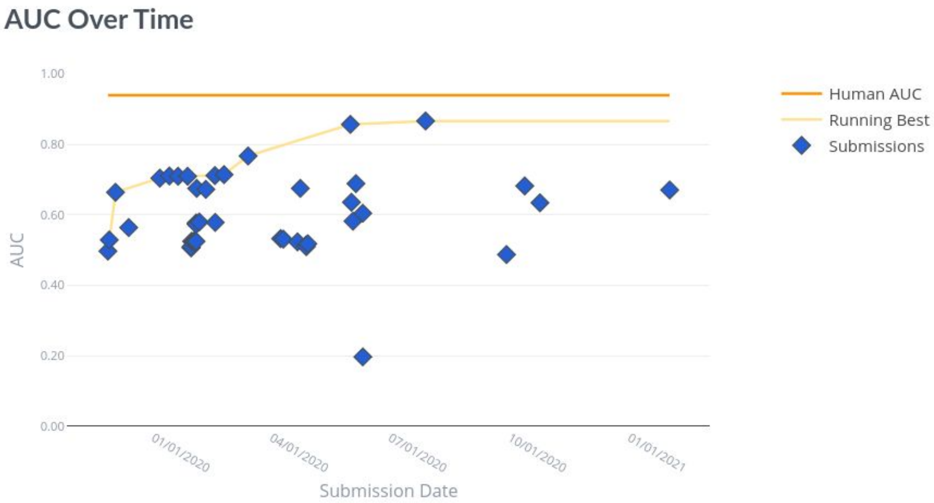
- Was considered a hard task for years
- Until recently, models' performance oscillated near random
- But now things looks different





The Winograd Schema

Did the muppets solve it?



The Winograd Schema

Did the muppets solve it?

No!

The Winograd Schema: Issue #1

Models perform better than random even with partial information

The Winograd Schema: Biases

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

The Winograd Schema: Biases

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

No-Candidates

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

The Winograd Schema: Biases

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

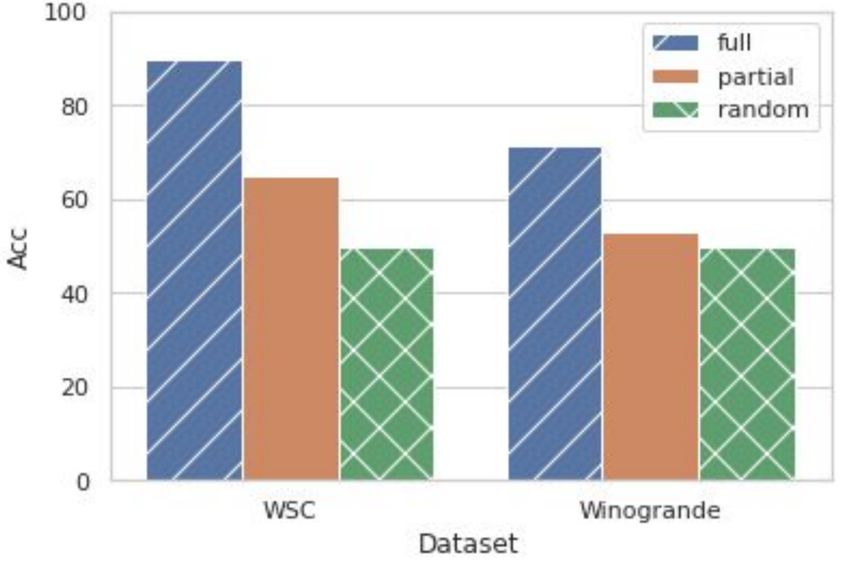
The Winograd Schema: Biases

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

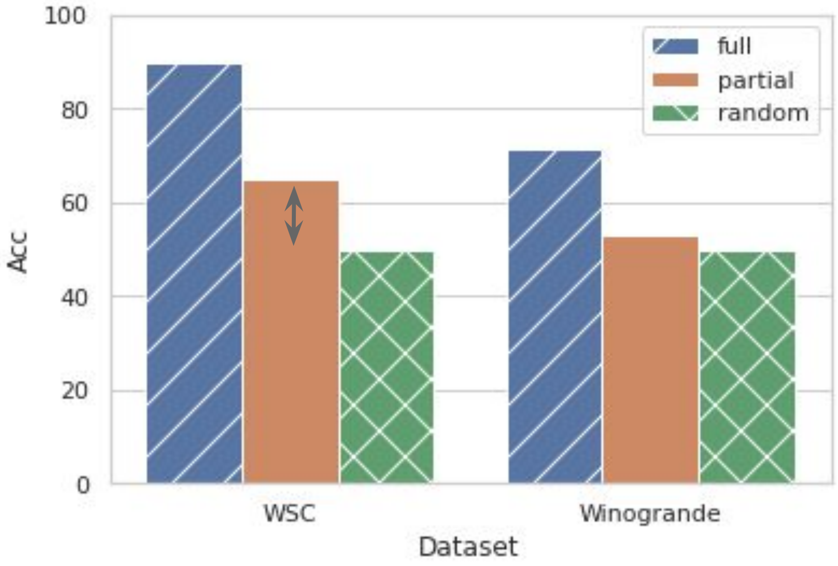
Partial-Sentence

- because it is too large.

The Winograd Schema: Biases

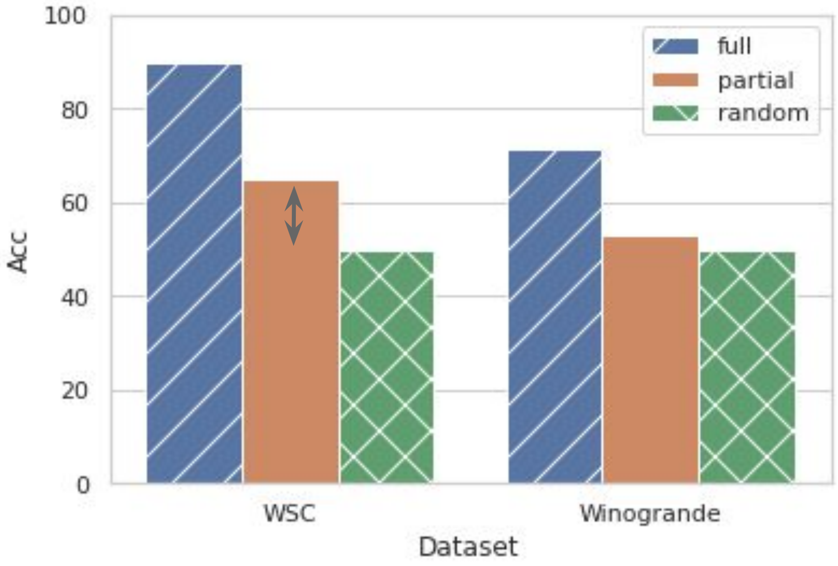


The Winograd Schema: Biases



The Winograd Schema: Biases

The gap from random shrinks!



The Winograd Schema: Evaluation

Current evaluation is sub-optimal

The Winograd Schema: Evaluation

Every winograd example constitutes of paired sentences:

- 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.

The Winograd Schema: Evaluation

Every winograd example constitutes of paired sentences:

- 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.

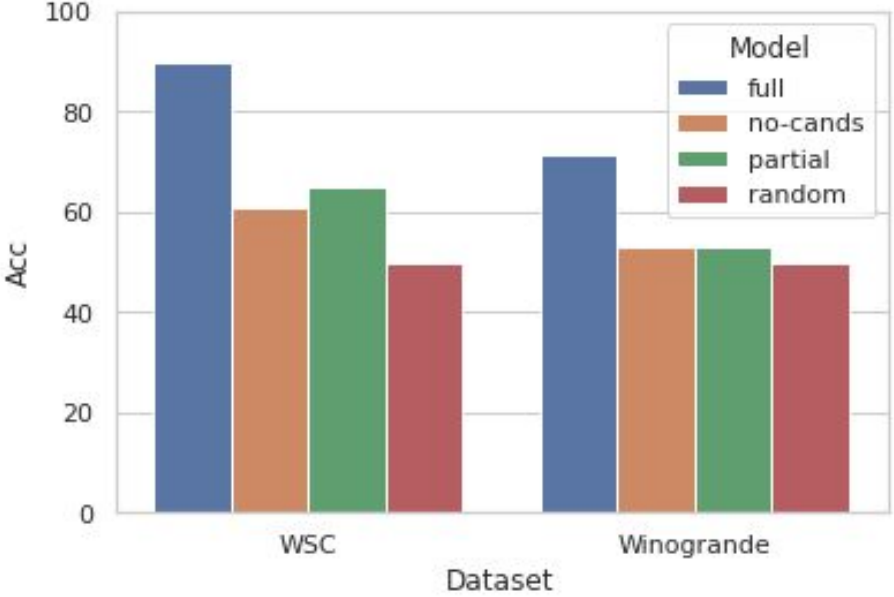
Succeeding on one may be due to randomness or some correlation

The Winograd Schema: Evaluation

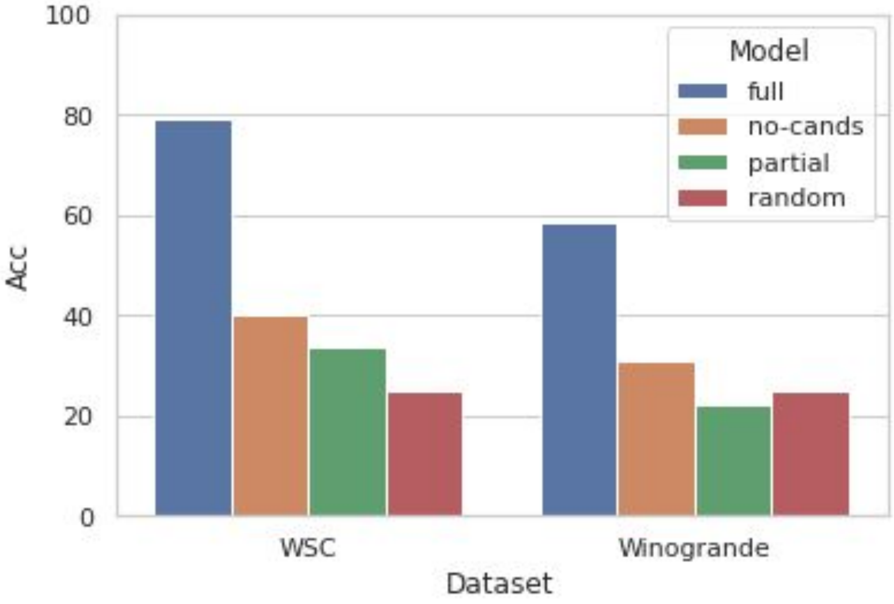
Instead, we require correct predictions on each pair

- A more robust evaluation
- Reduces the risk of “giving away” points to biased examples

The Winograd Schema: Evaluation



The Winograd Schema: Evaluation

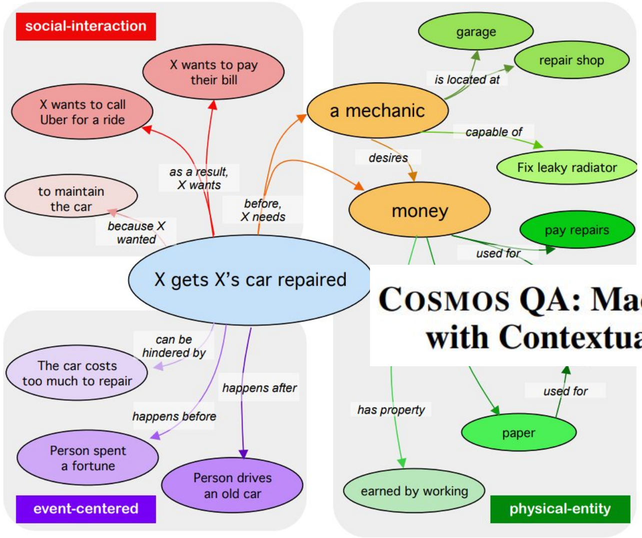


The Winograd Schema: Issue #2

Training on commonsense reasoning is **futile**

The Winograd Schema: Training

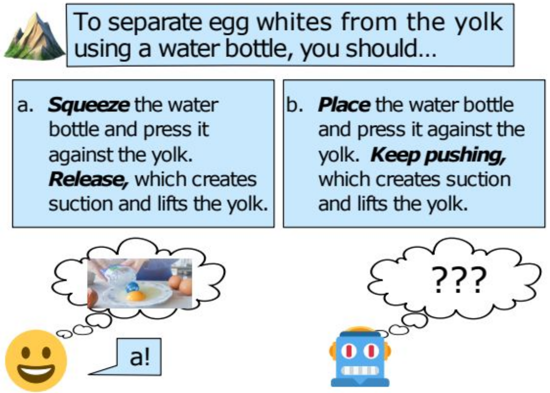
- Commonsense space is huge



COSMOS QA: Machine Reading Comprehension with Contextual Commonsense Reasoning

***HellaSwag*: Can a Machine *Really* Finish Your Sentence?**

PIQA: Reasoning about Physical Commonsense in Natural Language



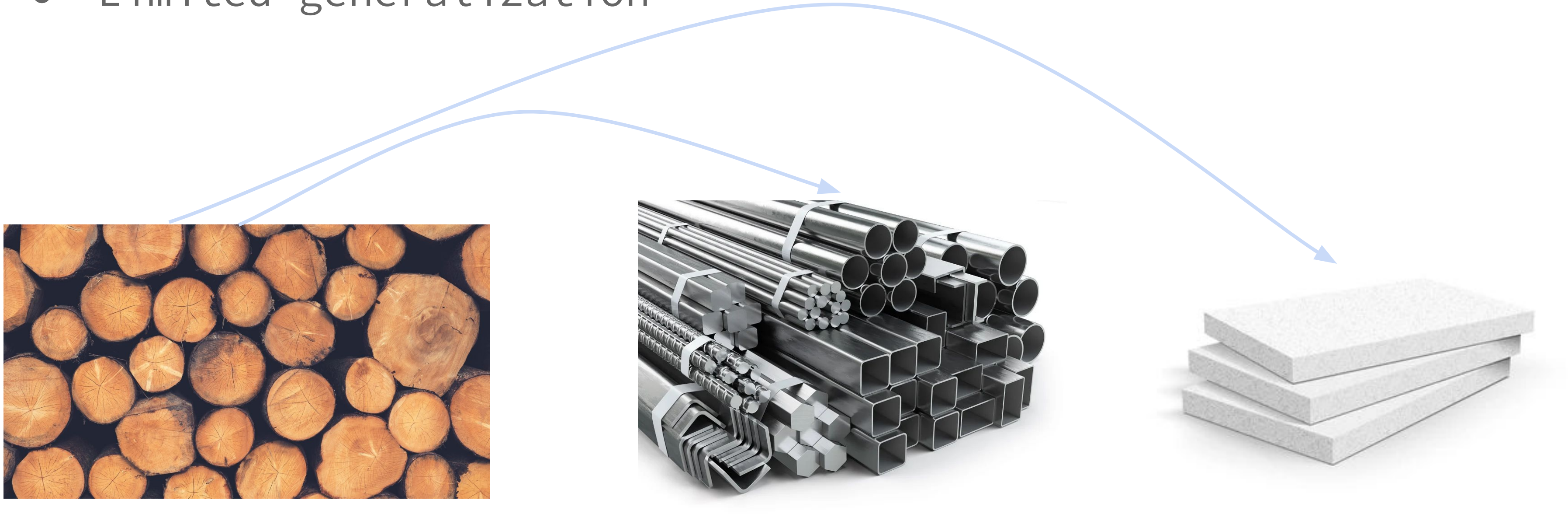
The Winograd Schema: Training

- Commonsense space is huge

Are supervised datasets with 10K, 100K enough?

The Winograd Schema: Training

- Commonsense space is huge
- Limited generalization



The Winograd Schema: Training

Instead:

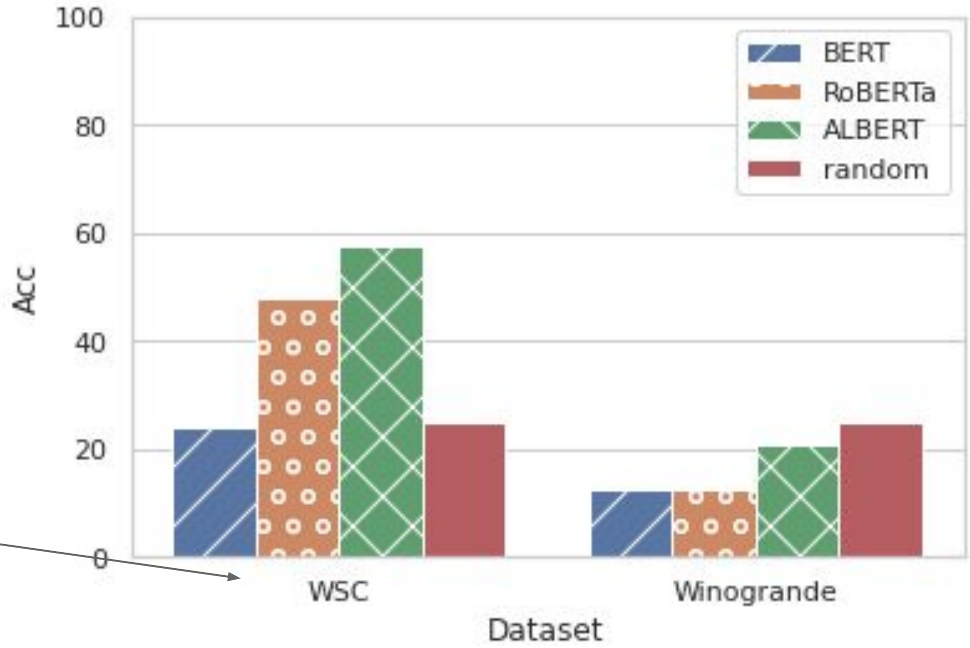
- Evaluate in a zero-shot / few-shot setting

Using Masked-Language Models:

- The trophy doesn't fit into the brown suitcase because the trophy is too ___.
  (large/small)

The Winograd Schema: Training

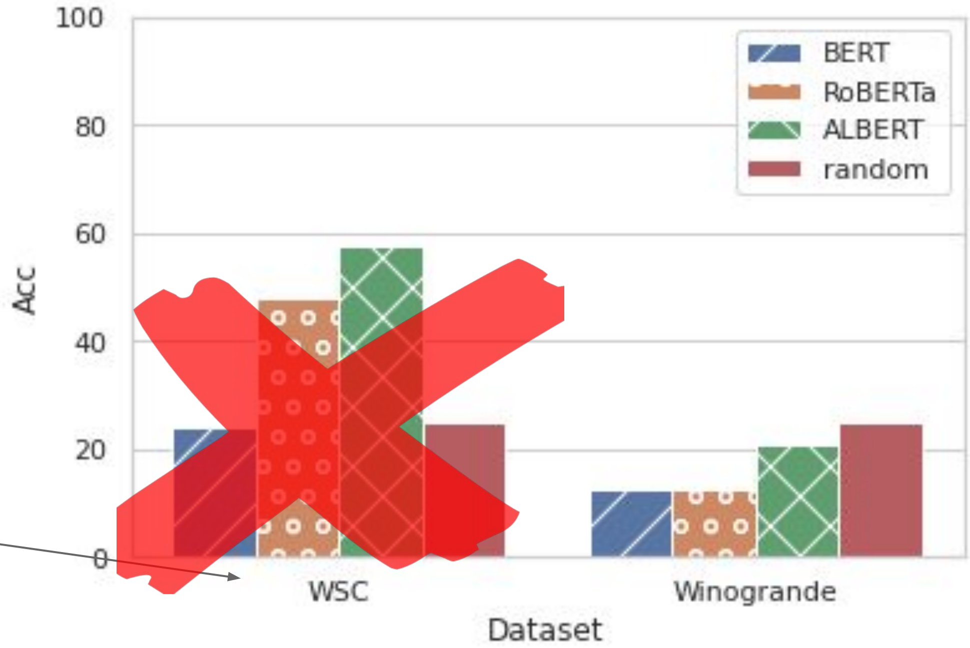
- Zero-Shot performance:



Biased,
Small

The Winograd Schema: Training

- Zero-Shot performance:

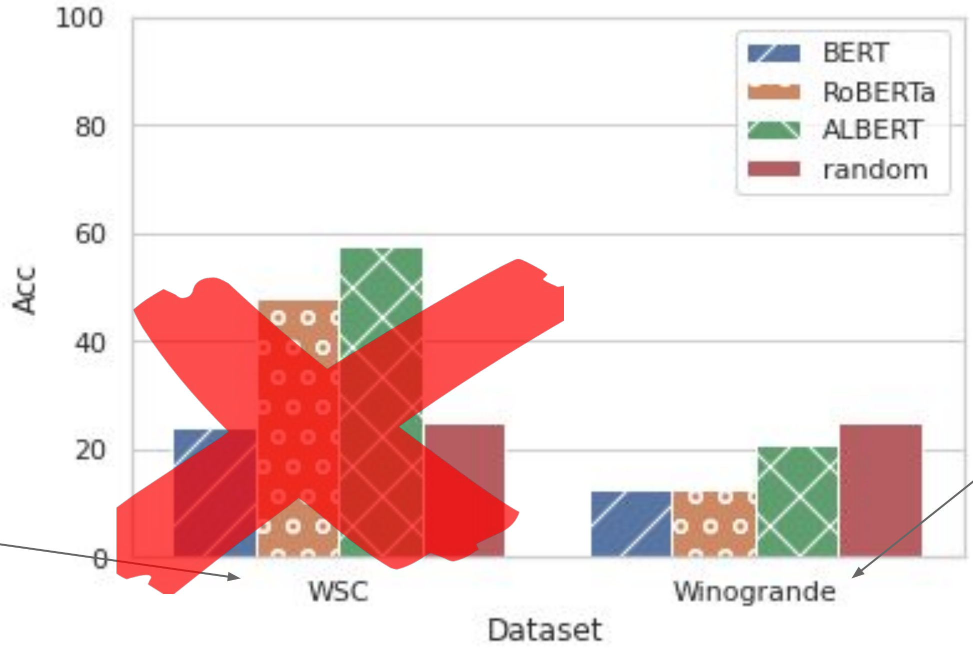


Biased, Small

The Winograd Schema: Training

- Zero-Shot performance:

Biased,
Small



Much less biased,
Large

The Winograd Schema

Although leaderboards seem to be solved, we’re still far from human agreement

Case Study II

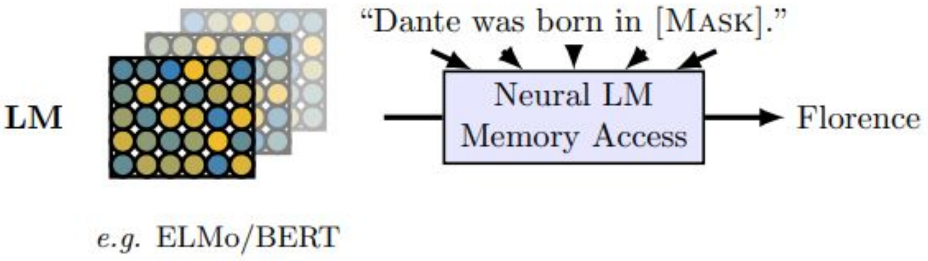
Consistency & Knowledge

Or, when iCloud was created both by Google and Sony

Consistency & Knowledge

- Language Models are trained over large text corpora
- As a by product, they retain factual knowledge

- These LMs are thought to provide good language understanding capabilities
- Thus, they should provide some language-based interface



Consistency & Knowledge

Are these models consistent to knowledge?

I.e. given two paraphrases, will the answer remain the same?

- "*Seinfeld* was originally aired on ___"
- "*Seinfeld* was premiered on ___"

Consistency & Knowledge: ParaRel 🤘

# Relations	38
# Patterns	328
Min # patterns	2
Max # patterns	20
Avg # patterns	8.63
Avg syntax	4.74
Avg lexical	6.03

Consistency & Knowledge

Standard
Task
Performance

Model	Accuracy	Consistency
majority	23.1+-21.0	100.0+-0.0
BERT-base	45.8+-25.6	58.5+-24.2
BERT-large	48.1+-26.1	**61.1**+-23.0
BERT-large-wwm	**48.7**+-25.0	60.9+-24.2
RoBERTa-base	39.0+-22.8	52.1+-17.8
RoBERTa-large	43.2+-24.7	56.3+-20.4
ALBERT-base	29.8+-22.8	49.8+-20.1
ALBERT-xxlarge	41.7+-24.9	52.1+-22.4

Consistency
+
Accuracy

What did we achieve
And where Are We Headed?

NLP - Today

- Bigger Models
- Bigger Training Corpora
- Human Performance

NLP - Today

- Their increasing sizes allow them to memorize the internet
- Good representation + Big datasets = Human performance

NLP - Today

However!

- Reading the entire internet doesn’t make you smart
- LMs are merely large capacity, statistical models

NLP - The Future

- Better Evaluation
    - More than a single evaluation metric
    - Size and latency measurements

NLP - The Future

- Models that
  - Have commonsense knowledge
  - Can make causal inferences
  - Are Grounded in knowledge

Thanks for Listening

Yanai Elazar

@yanaiela

yanaiela.github.io