Winograd Schema Challenge

Data Circles Journal Club 5-25-22

Reverse Timeline

2017	Large Language Models: Transformers
2010s	Probabilistics Models: Neural Networks, RNN/GRU/LSTM
1990s	Statistical Models: n-gram co-occurrence
pre-90s	Rule-Based Systems

You shall know a word by the company it keeps

~ J.R. Firth, Linguist

Encoder-Decoder

Fine-tuned for Downstream Tasks

Many-to-Many (Seq2Seq)

- Machine translation
- Text summarization
- Question answering

Examples: T5, BART

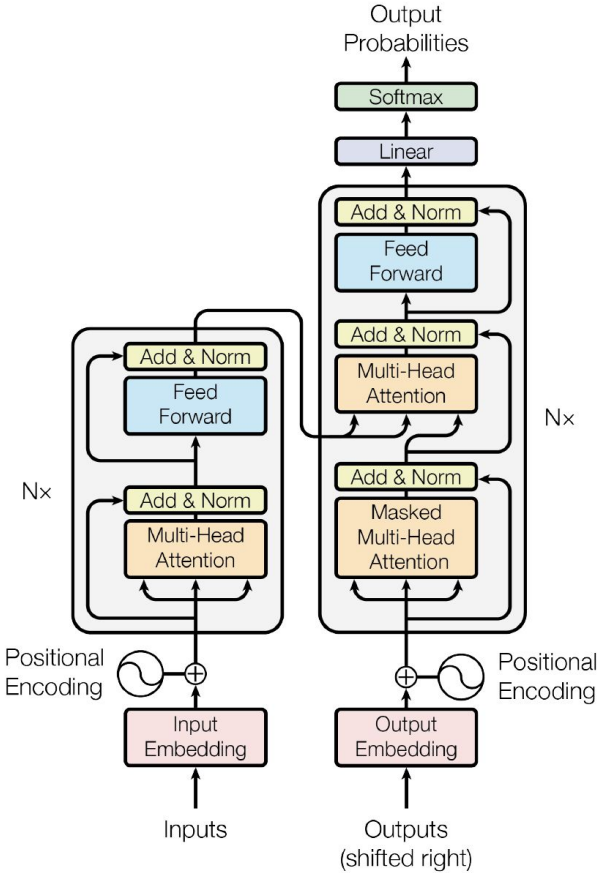


Figure 1: The Transformer - model architecture.

Encoder Only

Many-to-1

- Sentiment analysis
- Classification

Examples: BERT, RoBERTa, many variants

Masked (Cloze) Language Modeling

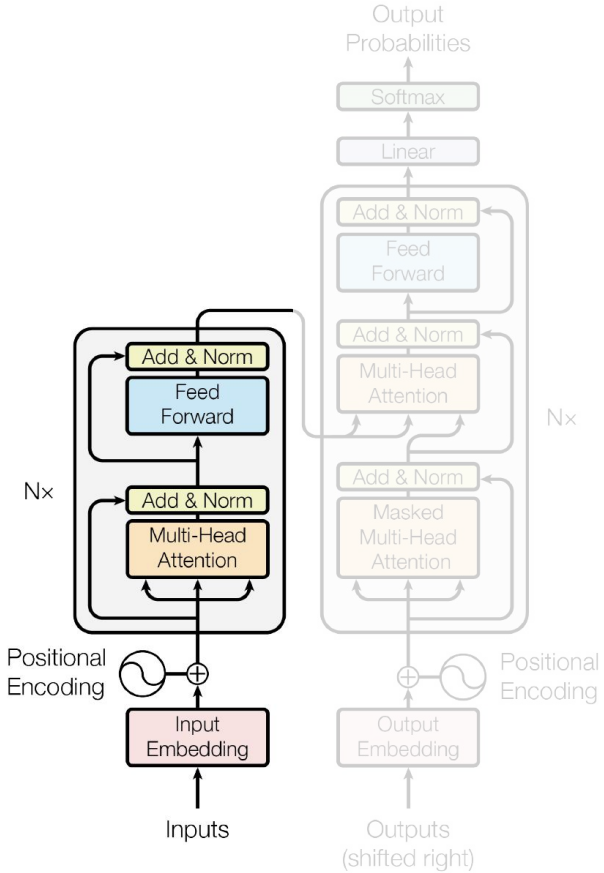


Figure 1: The Transformer - model architecture.

Decoder Only

1-to-Many

- Generation

Examples: GPT, GPT-2, GPT-3

Causal (Autoregressive) Language Modeling

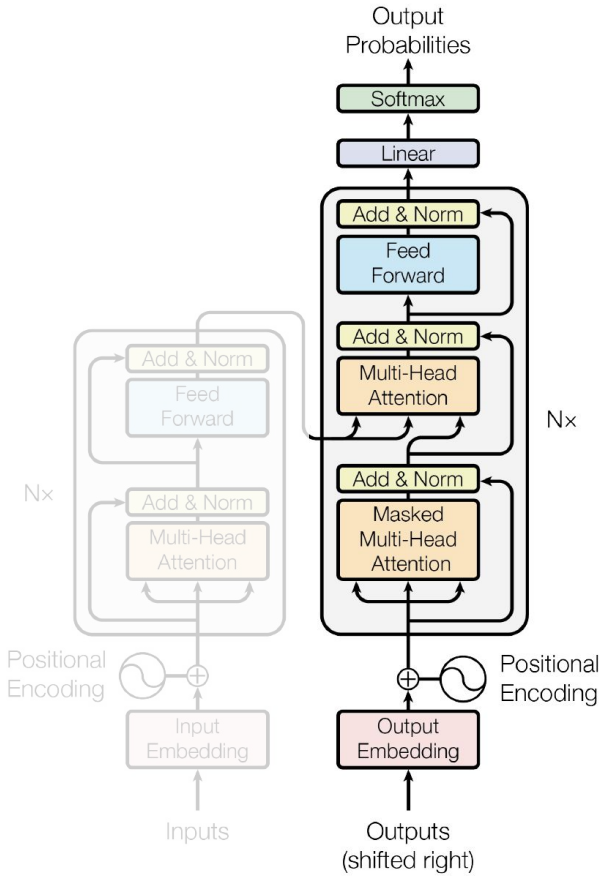
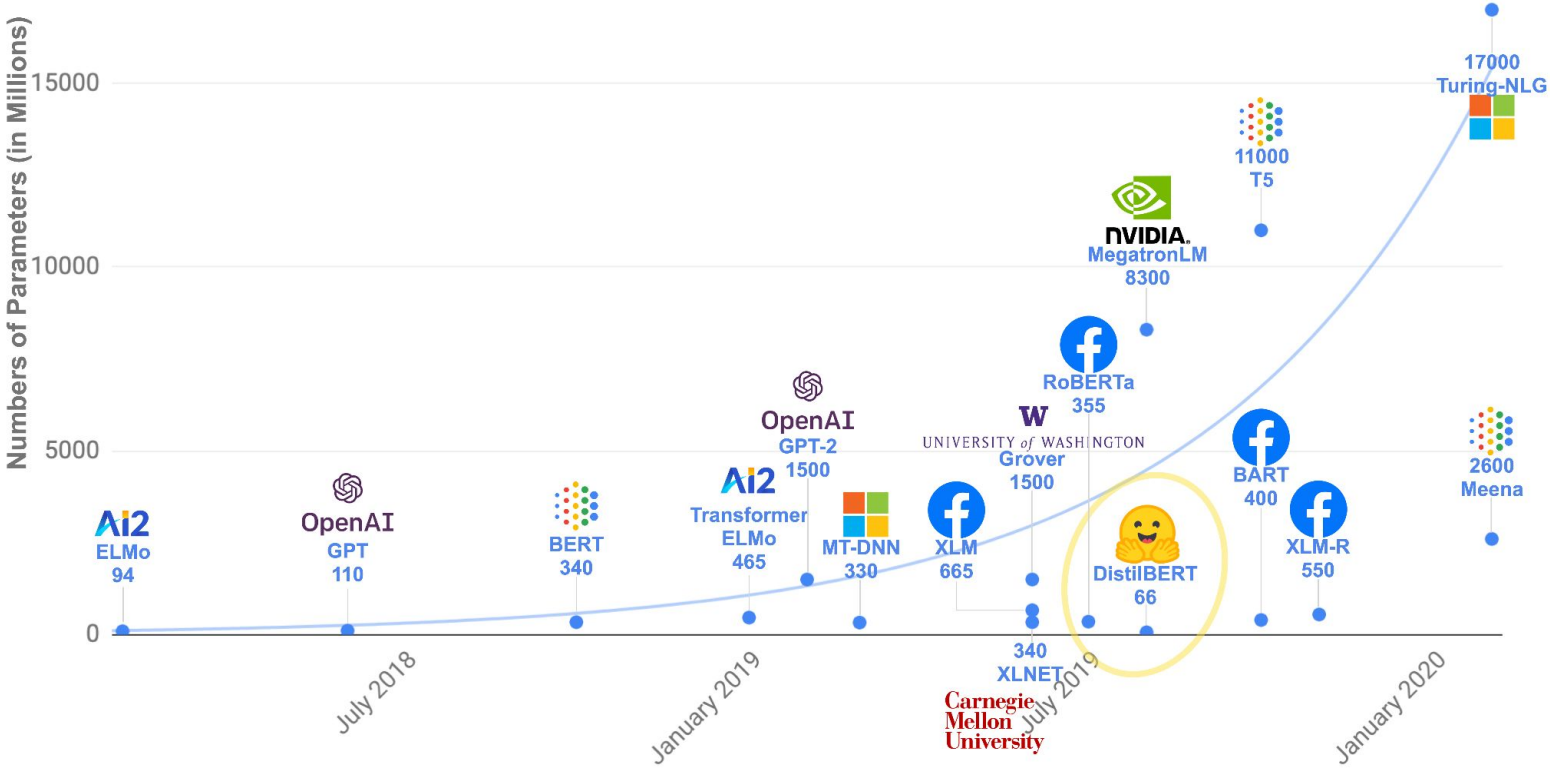
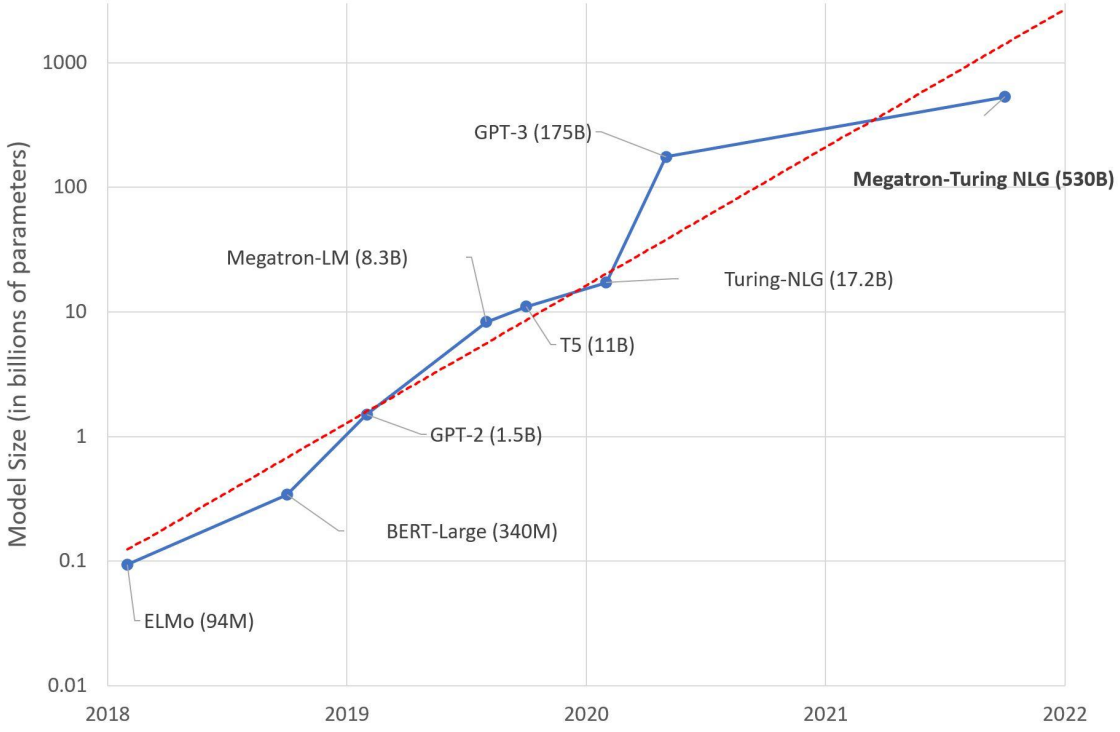


Figure 1: The Transformer - model architecture.

Large Language Models



Large Language Models



Reverse Timeline

2017 Large Language Models: Transformers

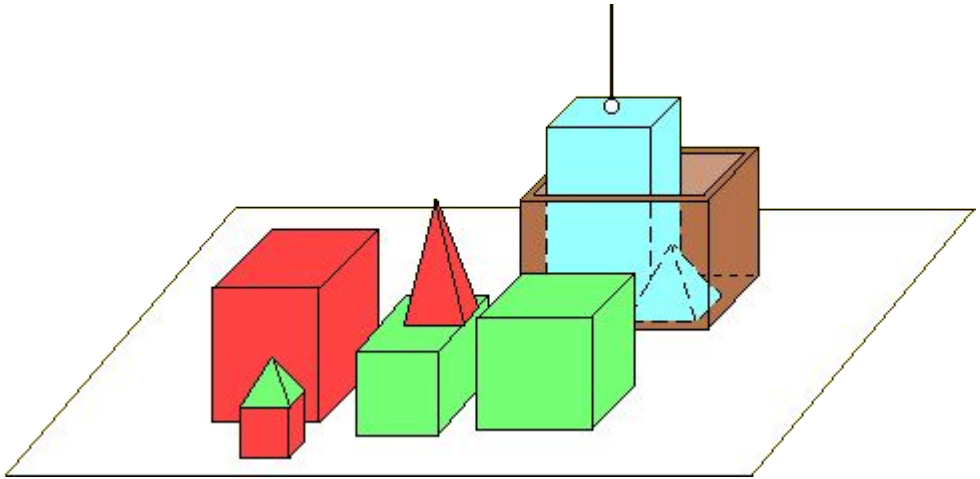
2010s Probabilistics Models: Neural Networks, RNN/GRU/LSTM

1990s Statistical Models: n-gram co-occurrence

pre-90s Rule-Based Systems

60-70s 1966 ELIZA, 1970 SHRDLU

1950 Turing Test



The Most Human Human

BRIGHTON, ENGLAND, SEPTEMBER 2009. I wake up in a hotel room 5,000 miles from my home in Seattle...

In two hours, I will sit down at a computer and have a series of five-minute instant-message chats with several strangers.

At the other end will be a psychologist, a linguist, a computer scientist, and the host of a popular British technology show.

Together they form a judging panel, evaluating my ability to do one of the strangest things I've ever been asked to do.

I must convince them that I'm human. Fortunately, I ***am*** *human; unfortunately, it's not clear how much that will help.*

The Winograd Schema Challenge

Hector J. Levesque
Dept. of Computer Science
University of Toronto
Toronto, Ontario
Canada M5S 3A6
hector@cs.toronto.edu

Ernest Davis
Dept. of Computer Science
New York University
New York, NY 10012
davise@cs.nyu.edu

Leora Morgenstern
S.A.I.C
Arlington, VA 22203
leora.morgenstern@saic.com

Abstract

In this paper, we present an alternative to the Turing Test that has some conceptual and practical advantages. A Winograd schema is a pair of sentences that differ only in one or two words and that contain a referential ambiguity that is resolved in opposite directions in the two sentences. We have compiled a collection of Winograd schemas, designed so that the correct answer is obvious to the human reader, but cannot easily be found using selectional restrictions or statistical techniques over text corpora. A contestant in the Winograd Schema Challenge is presented with a collection of one sentence from each pair, and required to achieve human-level accuracy in choosing the correct disambiguation.

ing the presence of thinking (or understanding, or intelligence, or whatever appropriate mental attribute), we assume that typed English text, despite its limitations, will be a rich enough medium.

2 The trouble with Turing

The Turing Test does have some troubling aspects, however. First, note the central role of *deception*. Consider the case of a future intelligent machine trying to pass the test. It must converse with an interrogator and not just show its stuff, but fool her into thinking she is dealing with a *person*. This is just a game, of course, so it's not really lying. But to imitate a person well without being evasive, the machine will need to assume a false identity (to answer "How tall are you?"

RTE – *Recognizing Textual Entailment*

yes-no questions concerning whether sentence A, called the **text**, entails another B, called the **hypothesis**.

A: Time Warner is the world's largest media and internet company.
B: Time Warner is the world's largest company.

A: Norway's most famous painting, "The Scream" by Edvard Munch, was recovered Saturday.
B: Edvard Munch painted "The Scream."

CoPA – *Choice of Plausible Alternatives*

Premise: I knocked on my neighbor’s door.
What happened as a **result**?
Alternative 1: My neighbor invited me in.
Alternative 2: My neighbor left his house.

Premise: The man fell unconscious.
What was the **cause** of this?
Alternative 1: The assailant struck the man in the head.
Alternative 2: The assailant took the man’s wallet.

Original Winograd Schema

The town council members refused to give the angry demonstrators a permit because they feared violence.

Who feared violence?
Answer 0: the town council members
Answer 1: the angry demonstrators

Original Winograd Schema

The town council members refused to give the angry demonstrators a permit because they **advocated** violence.

Who **advocated** violence?
Answer 0: the town council members
Answer 1: the angry demonstrators

Key Features

Pronoun resolution or pronoun disambiguation

2 noun phrases + 1 pronoun

Question what the pronoun refers to

- Answer 0 = 1st noun
- Answer 1 = 2nd noun

A **special word** can be replaced by **alternate word**
Still makes perfect sense but changes answer

More Examples

The trophy doesn't fit in the brown suitcase because it's too **big**.
What is too **big**?
Answer 0: the trophy
Answer 1: the suitcase

The trophy doesn't fit in the brown suitcase because it's too **small**.
What is too **small**?
Answer 0: the trophy
Answer 1: the suitcase

More Examples

Joan made sure to thank Susan for all the help she had **given**.
Who had **given** the help?
Answer 0: Joan
Answer 1: Susan

Joan made sure to thank Susan for all the help she had **received**.
Who had **received** the help?
Answer 0: Joan
Answer 1: Susan

Winograd Schema

273 pairs of sentences differing by only 1-2 words

Requires *knowledge and common-sense reasoning* to resolve

1. Easy for humans to understand
2. Not solvable by simple techniques
3. Google-proof: no obvious statistical patterns

Obvious to non-experts whether program fails or succeeds

Pitfalls – *too easy*

The women stopped taking the pills because they were **[pregnant/carcinogenic]**.

Which individuals were **[pregnant/carcinogenic]**?

Answer 0: the women
Answer 1: the pills

Pitfalls – *too ambiguous*

Frank was **[jealous/pleased]** when Bill said that he was the winner of the competition.

Who was the winner?

Answer 0: Frank
Answer 1: Bill

GLUE: A MULTI-TASK BENCHMARK AND ANALYSIS PLATFORM FOR NATURAL LANGUAGE UNDERSTANDING

Alex Wang[1], **Amanpreet Singh**[1], **Julian Michael**[2], **Felix Hill**[3],
Omer Levy[2] **& Samuel R. Bowman**[1]
[1]Courant Institute of Mathematical Sciences, New York University
[2]Paul G. Allen School of Computer Science & Engineering, University of Washington
[3]DeepMind
{alexwang,amanpreet,bowman}@nyu.edu
{julianjm,omerlevy}@cs.washington.edu
felixhill@google.com

ABSTRACT

For natural language understanding (NLU) technology to be maximally useful, it must be able to process language in a way that is not exclusive to a single task, genre, or dataset. In pursuit of this objective, we introduce the General Language Understanding Evaluation (GLUE) benchmark, a collection of tools for evaluating the performance of models across a diverse set of existing NLU tasks. By including tasks with limited training data, GLUE is designed to favor and encourage models that share general linguistic knowledge across tasks. GLUE also includes a hand-crafted diagnostic test suite that enables detailed linguistic analysis of models. We evaluate baselines based on current methods for transfer and representation learning and find that multi-task training on all tasks performs better than training a separate model per task. However, the low absolute performance of our best model indicates the need for improved general NLU systems.

GLUE Evaluation Tasks

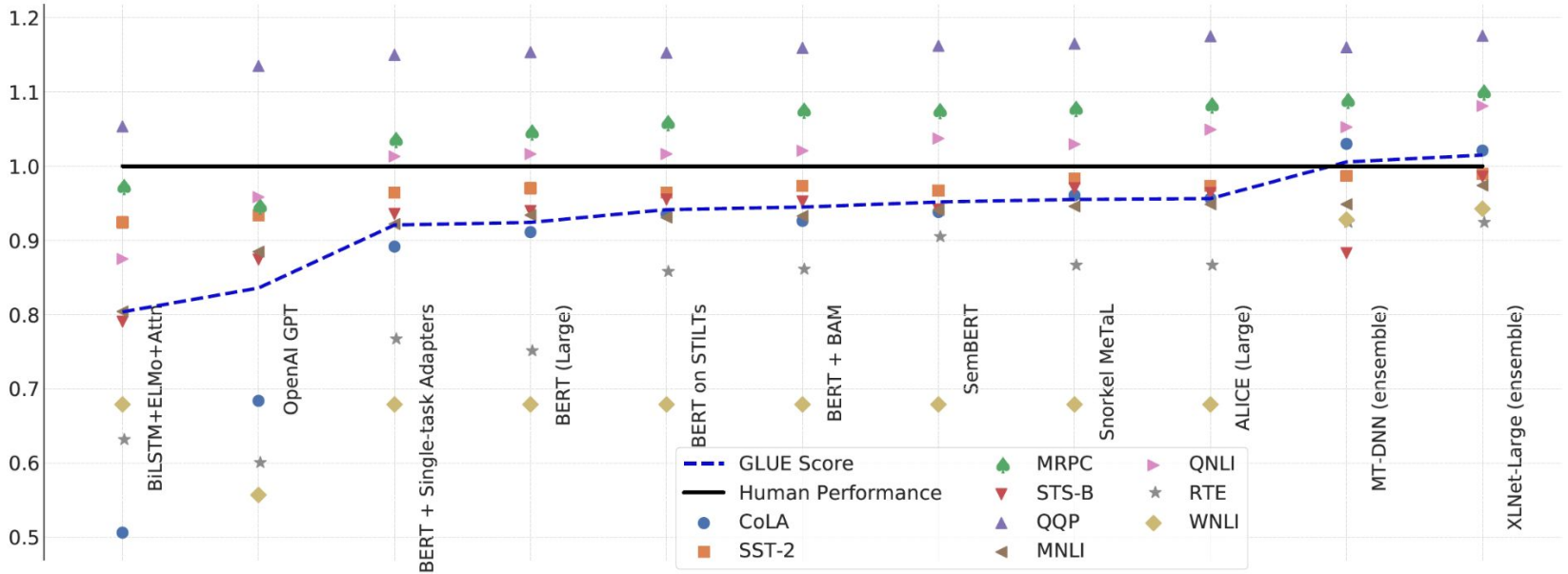
Single Sentence: CoLA, SST-2

Similarity & Paraphrase: MRPC, STS-B, QQP

Inference

- MNLI
- QNLI
- RTE
- WNLI

GLUE Results



The Defeat of the Winograd Schema Challenge

Vid Kocijan[a], Ernest Davis[b], Thomas Lukasiewicz[a,c], Gary Marcus[d], Leora Morgenstern[e]

[a] *University of Oxford, Department of Computer Science, Oxford, OX1 3QD, UK*

[b] *New York University, Department of Computer Science, 251 Mercer St, NY 10012, United States*

[c] *Alan Turing Institute, 96 Euston Rd, London NW1 2DB*

[d] *Robust AI, 380 Portage Avenue Palo Alto, CA 94306 United States*

[e] *Palo Alto Research Center, 3333 Coyote Hill Rd, Palo Alto, CA 94304, United States*

Abstract

The Winograd Schema Challenge—a set of twin sentences involving pronoun reference disambiguation that seem to require the use of commonsense knowledge—was proposed by Hector Levesque in 2011. By 2019, a number of AI systems, based on large pre-trained transformer-based language models and fine-tuned on these kinds of problems, achieved better than 90% accuracy. In this paper, we review the history of the Winograd Schema Challenge and assess its significance.

Keywords: Commonsense Reasoning, Winograd Schema Challenge

Winograd Schema Timeline

1972	Winograd's PhD thesis introduces original example
2011	Levesque proposes Winograd Schema Challenge Creates initial corpus
2016	Winograd Schema Challenge run IJCAI-17. No systems better than chance
2018	WNLI incorporated in GLUE benchmarks. BERT-based systems do no better than most-frequent-class guessing
2019	RoBERTa 89.0%, WinoGrande 90.1%

Survey Related Datasets

Winograd Natural Language Inference Dataset (WNLI, SuperGLUE WSC)

Definite Pronoun Resolution (DPR)

Pronoun Disambiguation Problem (PDP)

WinoGender & WinoBias

WinoGrande / WinoFlexi / Winventor

WinoWhy & WinoLogic

WinoMT & Wino-X

Languages: French, Portuguese, Japanese, Chinese, Hindi, Slovene, Hebrew

PDP Pronoun Disambiguation Problem

Found in wild. Easier to collect

Mrs. March gave the mother tea and gruel, while she dressed the little baby as tenderly as if it had been her own

Mama came over and sat down beside Sarah. Gently she stroked her hair and let the child weep.

The Scarecrow seized the oilcan from Dorothy's basket and oiled the Woodman's jaws, so that after a few moments he could talk as well as before.

WinoGender & WinoBias

Diagnostic to measure bias

The surgeon operated on the child with great care;
[his/her] [tumor/affection] had grown over time.

The farmer knows the editor because **[he/she]**
[is really famous/likes the book].

The accountant met the janitor and wished **[her/him]** well.

WinoMT & WinoX

Diagnostic to measure cross-lingual bias

WinoGrande

Crowdsourced + adversarial filtering. SuperGLUE

Spatial relations, perception
The sack of potatoes had been placed **[above/below]** the bag of flour, so it had to be moved first
There is a pillar between me and the stage, and I can't see **[around it/it]**

Social behaviors/interactions/emotions
Bob paid for Charlie's college education, but now Charlie acts as though it never happened. He is very **[hurt/ungrateful].**
Alice tried frantically to stop her daughter from **[chatting/barking]** at the party, leaving us to wonder why she was behaving so strangely.

WinoGrande

Problems

The doctor diagnosed Justin with bipolar and Robert with anxiety. **[Justin/Robert]** had terrible nerves recently.

The waiter could not cover the round tables with the square tablecloths because the **[tables/tablecloths]** were square.

George opted for both of them to use a knife instead of a gun in the duel because the **[knife/gun]** could partially injure them.

WinoFlexi & Winventor

Crowdsourced + automated candidate generation

As **Frederick** was rather distant to **his family**, **Eleanor** had a great influence on the raising and education of **Frederick's children**, and she therefore played an important role in the House of Hapsburg's rise to prominence.

Who therefore played an important role in the House of Hapsburg's rise to prominence?

WinoWhy & WinoLogic

Explanations of WSC273

Human Evaluations

Dataset	% correct
Experimental dataset of 32 schemas	93.0%
143 schemas from WSC273	92.1%
66 texts with 108 PDPs	90.1%
89 unpublished Winograd schemas	92%
Unpublished collection of 86 texts with 101 PDPs	96%
WNLI	96.1%
WinoGender	94.9%
WSC273	86.5%
WinoGrande	94%

Methods

Rule-Based Systems

Neural Networks

Large Language Models

	Language Model	fine-tuning or external data
Trinh and Le (2018)	custom LSTM	–
Radford et al. (2019)	GPT-2	–
Klein and Nabi (2019)	BERT	–
Prakash et al. (2019)	custom LSTM	internet querying
Kocijan et al. (2019b)	BERT	MaskedWiki, DPR
Kocijan et al. (2019a)	BERT	WikiCREM, DPR, GAP
Ruan et al. (2019)	BERT	DPR
He et al. (2019)	BERT	DPR
Ye et al. (2019)	BERT	ConceptNet, DPR
Sakaguchi et al. (2020)	RoBERTa	WinoGrande
Melo et al. (2020)	custom LSTM	–
Brown et al. (2020)	GPT-3	–
Yang et al. (2020)	RoBERTa	WinoGrande and generated data
Lin et al. (2020)	T5 (3B)	WinoGrande
Khashabi et al. (2020)	T5	WinoGrande and QA tasks
Lourie et al. (2021b)	T5	WinoGrande and RAINBOW

Table 4: Resources used by different language-model based approaches, ordered by the time of publication. With time, ever larger language models and more additional fine-tuning data was used.

Problems

Complaint small initial dataset. Not meant for training and tuning.

Solving pronoun resolution not *commonsense reasoning* or *intelligence*.

1. Lax evaluation criteria.
   Difficulties manually creating high-quality WSC.
2. Artifacts in the datasets that remain
3. Leakage from large training data

Pronoun disambiguation** of a model that has been fine-tuned to that task is **not** at all a reliable measure of the degree to which the model has learned **commonsense knowledge** broadly, or even to which it has learned the commonsense knowledge needed for **language understanding.

Discussion

Personal experience with language models or NLP?

What’s the best way to evaluate intelligence and common sense?

How do humans develop attention and common sense?

Thoughts whether brain uses rule-based or probabilistic algorithms?