Do language models learn *typicality* judgments from text?

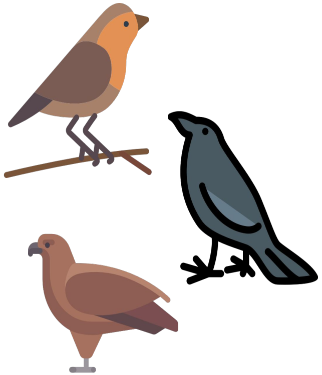
Kanishka Misra🦅, Allyson Ettinger🦉, Julia Taylor Rayz🦅
🦅 Purdue University
🦉 University of Chicago

CogSci 2021

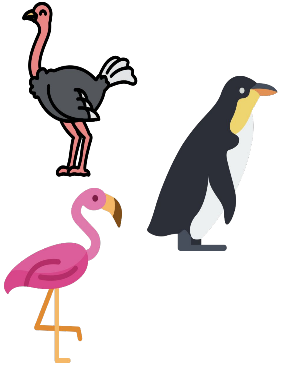
Typicality

Some items of a category are more representative members than others.

Typical Birds



Atypical Birds



Rosch et al., 1975

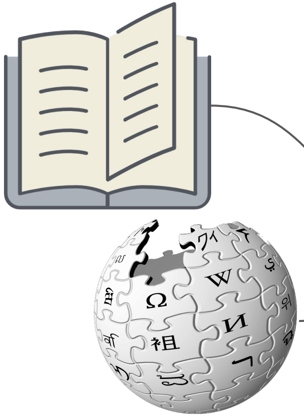
Member	Goodness of example: Rank	Goodness of example: Specific score
robin	1	1.02
sparrow	2	1.18
bluejay	3	1.29
bluebird	4	1.31
canary	5	1.42
blackbird	6	1.43
dove	7	1.46
lark	8	1.47
swallow	9	1.52
parakeet	10	1.53
oriole	11	1.61
mockingbird	12	1.62
redbird	13.5	1.64
wren	13.5	1.64
finch	15	1.66
starling	16	1.72
cardinal	17.5	1.75
eagle	17.5	1.75
hummingbird	19	1.76
seagull	20	1.77
woodpecker	21	1.78
pigeon	22	1.81
thrush	23	1.89
falcon	24	1.96
crow	25	1.97
hawk	26	1.99
raven	27	2.01

Ubiquity of *Typicality*

Typicality affects:

- **Taxonomic sentence verification** (Rips et al., 1973; Rosch, 1973)
- **Exemplar production order** (Rosch et al., 1976)
- **Concept acquisition** (Rosch et al., 1976)
- **Category-based Induction** (Osherson et al., 1990)
- … many more!

Learning from Language using Language Models



Language Models

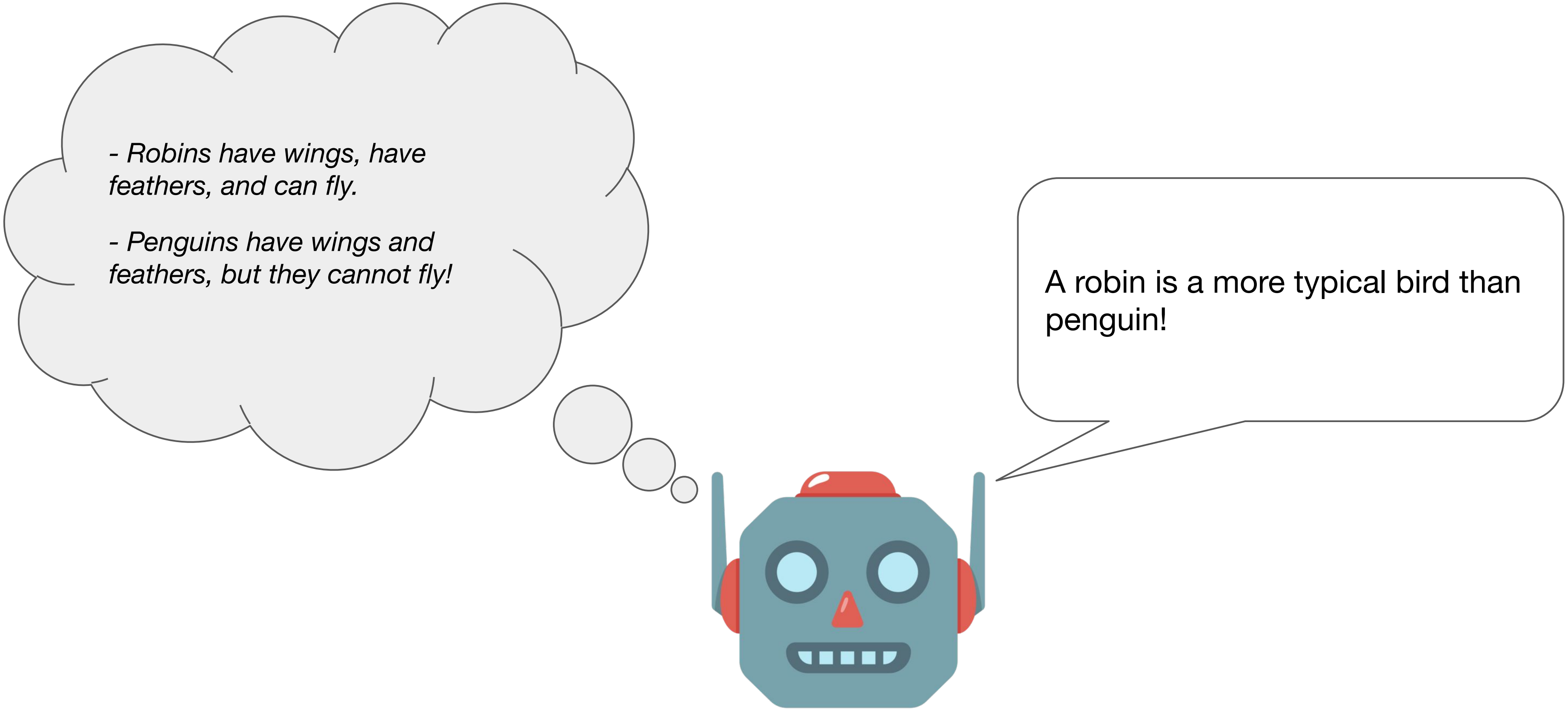
P(word | context)

Factual Knowledge (Petroni et al., 2019):
Paris is the capital of France.

Categorical Knowledge (Ettinger, 2020):
A robin is a bird.

Concept Knowledge (Weir et al., 2020):
A bird has wings, has feathers, can fly

TYPICALITY


- *Robins have wings, have feathers, and can fly.*
- *Penguins have wings and feathers, but they cannot fly!*
A robin is a more typical bird than penguin!

Taxonomic Sentence Verification

Phenomenon: Typicality promotes faster sentence verification.

Most typical **Least typical**

RT("*A **robin** is a bird.*") < ….. < RT("*An **eagle** is a bird.*") < ….. < RT("*An **ostrich** is a bird.*")

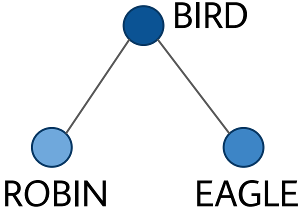
Rips, Shoben, and Smith, 1973;
Rosch, 1973

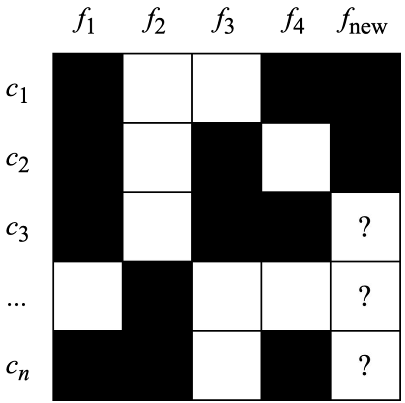
Category-based Induction

Inductive Reasoning: A premise-conclusion setup where the conclusion does not *necessarily* follow from the premise.

Premise — Robins have the T9 Hormone.
Eagles have the T9 Hormone.

Conclusion — All **birds** have the T9 Hormone.





Rips, 1975; Osherson et al., 1990; Kemp and Jerns, 2014; Feeney and Heit, 2007

Category-based Induction

Phenomenon: Subjects are more likely to generalize new information about a member *m* to the entire category when *m* is typical -- as opposed to atypical -- to the category.

Robins have property *P*.	Penguins have property *P*.
All birds have property *P*.	All birds have property *P*.

Property P = *blank*, i.e., The agent has minimal information about the property. E.g. *has sesamoid bones, has the T9 Hormone, loves onions.*

Osherson et al., 1990

Probing for *Typicality*

1) **Taxonomic Sentence Verification (Rips et al., 1973; Rosch et al., 1973)**

A robin is a bird. vs. *A penguin is a bird.*

2) **Category-based Induction (Osherson et al., 1990)**

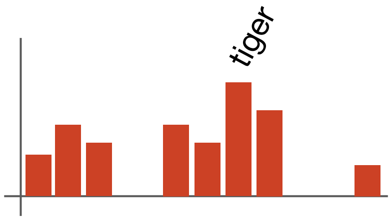
Robins can dax. $\longrightarrow$ *All birds can dax.*

vs.

Penguins can dax. $\longrightarrow$ *All birds can dax.*

Models Studied

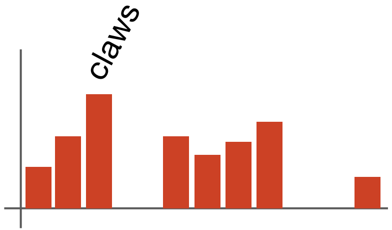
Masked Language Models



Bidirectional
Transformer

A [MASK] has claws.

Incremental Language Models



Unidirectional
Transformer

A tiger has

Models Studied

Masked Language Models

- 3 x **BERT (Devlin et al., 2019)**
- 3 x **RoBERTa (Liu et al., 2019)**
- 4 x **ALBERT (Lan et al., 2019)**
- 3 x **ELECTRA (Clark et al., 2020)**

Incremental Language Models

- 1 x **GPT (Radford et al., 2018)**
- 5 x **GPT2 (Radford et al., 2019)**

Baseline: 5-gram Language Model with KN smoothing trained on Wikipedia

Typicality Ratings

209 North American native English speakers tasked to rate goodness of example for 565 items across 10 categories.

Ratings from 1 (most typical) to 7 (least typical)

Category	*N*	Category	*N*
furniture	60	vegetable	56
tool	60	clothing	55
toy	60	bird	54
weapon	60	fruit	51
sport	59	vehicle	50

Member	Goodness of example: Rank	Goodness of example: Specific score
robin	1	1.02
sparrow	2	1.18
bluejay	3	1.29
bluebird	4	1.31
canary	5	1.42
blackbird	6	1.43
dove	7	1.46
lark	8	1.47
swallow	9	1.52
parakeet	10	1.53
oriole	11	1.61
mockingbird	12	1.62
redbird	13.5	1.64
wren	13.5	1.64
finch	15	1.66
starling	16	1.72
cardinal	17.5	1.75
eagle	17.5	1.75
hummingbird	19	1.76
seagull	20	1.77
woodpecker	21	1.78
pigeon	22	1.81
thrush	23	1.89
falcon	24	1.96
crow	25	1.97
hawk	26	1.99
raven	27	2.01

Stimuli

Taxonomic Sentence Verification

`[DET] [ITEM]` is `[DET] [CATEGORY]`.

N = 565

A robin is a bird.
An ostrich is a bird.
…
A hammer is a tool.

Category-based Induction

Blank properties: can dax, are vorpal, etc. (15-20 properties per item)

`[ITEM]`s [property-phrase].

All `[CATEGORY]`s [property-phrase].

N = 12,180

Robins can dax. All birds can dax.
Ostriches can dax. All birds can dax.
…
Hammers are slithy. All tools are slithy.

Measures

Taxonomic Sentence Verification

An ostrich is a bird.

$$TSV = \log P_{LM}(\textit{bird} \mid \textit{An ostrich is a})$$

Category-based Induction

Robins can fep. All birds can fep.

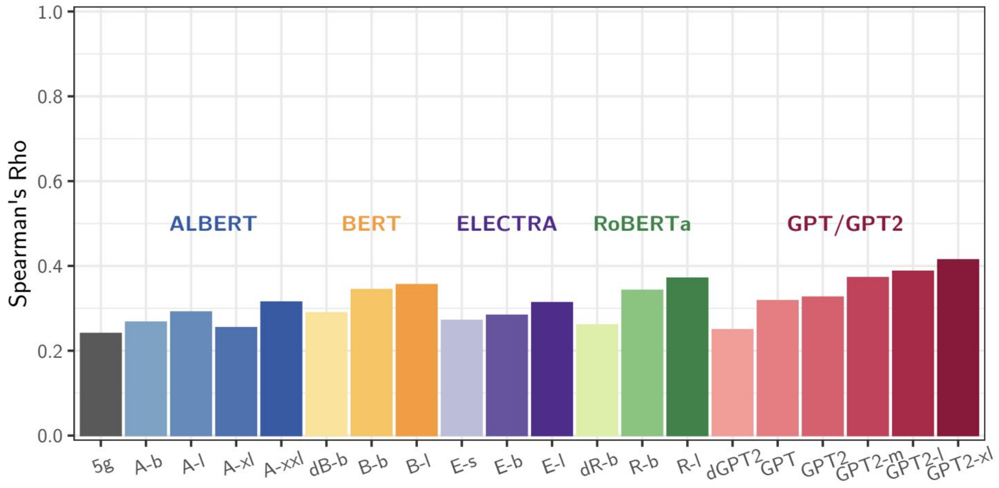
$$AS = \log P_{LM}(\textit{All birds can fep.} \mid \textit{Robins can fep.})$$

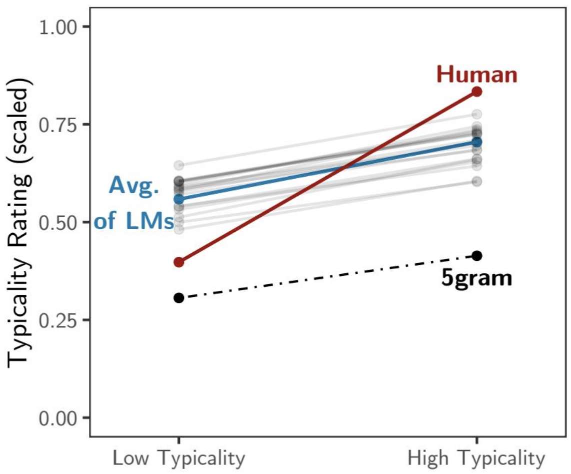
Results

Taxonomic Sentence Verification

$$\rho(\,\text{- } \textit{human rating} \mid \textit{TSV}\,)$$

$$\rho(\,\text{How typical a bird is robin?} \mid \log P_{LM}\,(\textit{bird} \mid \textit{A robin is a}))$$





Category-based Induction

$AS = \log P_{LM}($ *All birds can fep.* | ***Robins*** ***can fep.*** $)$

1) Premise order sensitivity (POS):

$\log P_{LM}($ *All birds can fep.* | ***Can fep robins.*** $)$

2) Taxonomic sensitivity (TS):

R^2 = 0.43

$\log P_{LM}($ *All birds can fep.* | ***Sofas can fep.*** $)$

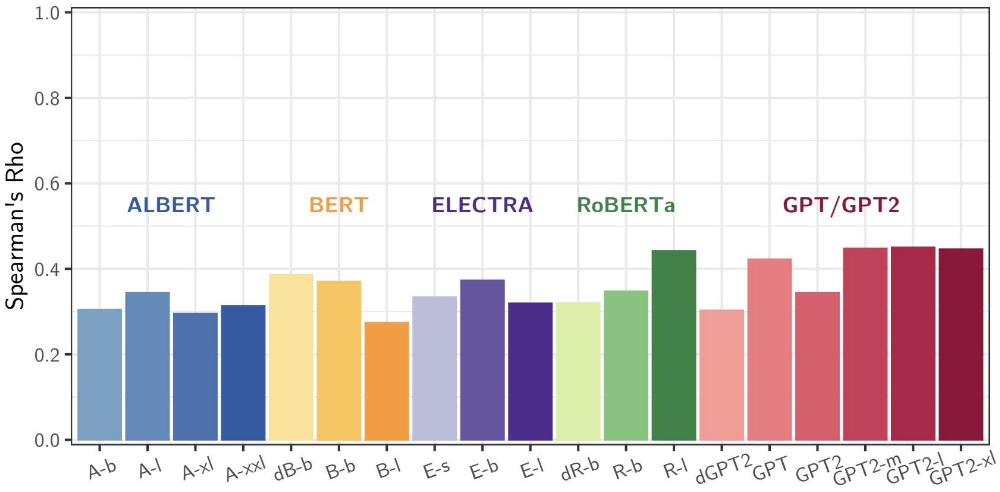
Regressing out confounds:

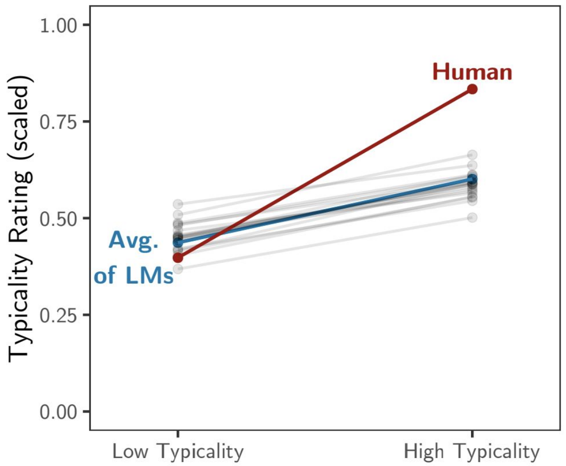
$$
\begin{aligned}
AS &= \beta_0 + \beta_1 \mathrm{TS} + \beta_2 \mathrm{POS} + \epsilon \\
AS' &= AS - \beta_1 \mathrm{TS} - \beta_2 \mathrm{POS} \\
&= \beta_0 + \epsilon \qquad\qquad \text{(Adjusted } AS\text{)}
\end{aligned}
$$

Category-based Induction

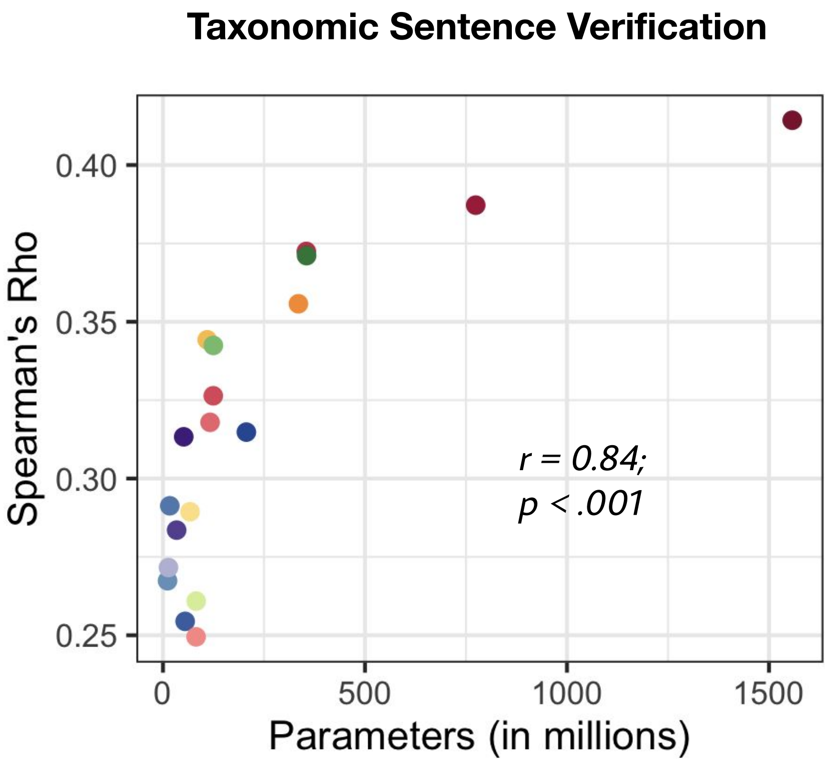
$$\rho(\,-\,\textit{human rating} \mid \textit{Adjusted AS}\,)$$

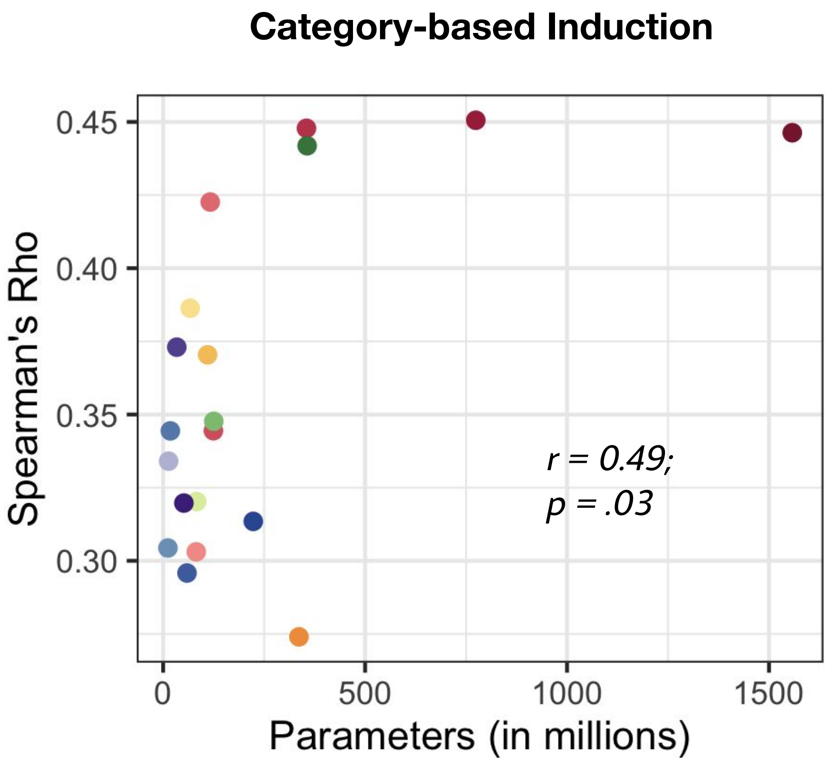
$$\rho(\text{ How typical a bird is robin? } \mid \log P_{LM}\,(\textit{All birds can dax.} \mid \textit{Robins can dax.}))$$





Relationship with # of Parameters





Takeaways & Speculations

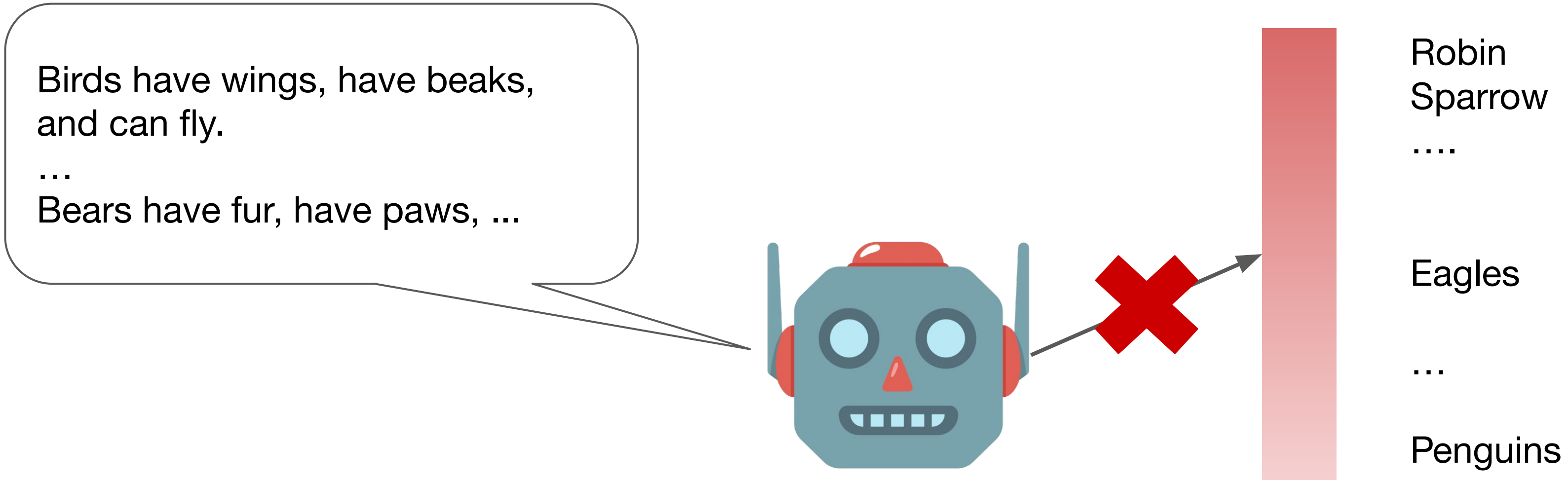
1. Word prediction capacities of LMs are moderately sensitive to human-elicited typicality ratings.

   As seen in:

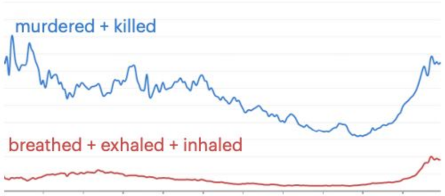
   a. **Attributing items to their category members.** (Scales with # of parameters.)

   b. **Making complex inductive inferences about categories when conditioned on new information about items.** (No clear relationship with # of parameters.)

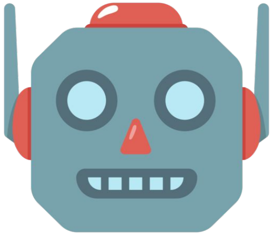
2. LMs show qualitatively similar patterns in distinguishing high and low typicality items, but are less extreme as compared to humans.


Birds have wings, have beaks, and can fly.
…
Bears have fur, have paws, ...
Robin
Sparrow
….
Eagles
…
Penguins

Takeaways & Speculations



Reporting Bias in Textual Corpora
(Gordon and Van Durme, 2013; Shwartz & Choi, 2020)



Children hear more about what is atypical than what is typical

Claire Bergey*
cbergey@uchicago.edu
Department of Psychology
University of Chicago

Benjamin C. Morris*
benmorris@uchicago.edu
Department of Psychology
University of Chicago

Daniel Yurovsky
yurovsky@cmu.edu
Department of Psychology
Carnegie Mellon University
University of Chicago

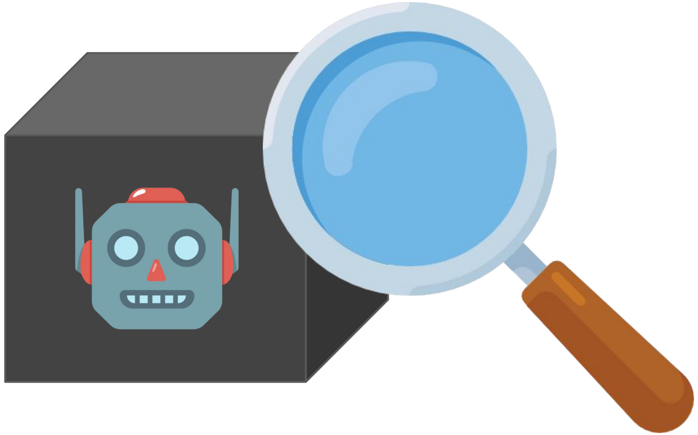
(CogSci 2020)

(Adjective-Noun Compounds)

Future Work

Train models:

- to correct for distorted frequencies of atypical items mentioned in text.
- informed by a more grounded source of knowledge.
- explicitly on features of categories and concepts (Rogers and McClelland, 2004; Bhatia and Ritchie, 2020)



Thanks!

"...if one compares different category members and does not find an effect of typicality, it suggests that there is something wrong--or at least unusual about--the experiment."

- Gregory Murphy (*The Big Book of Concepts*, 2004)

Code: `https://github.com/kanishkamisra/typicalityprobing`

kmisra@purdue.edu
aettinger@uchicago.edu
jtaylor1@purdue.edu

@kanishkamisra
@AllysonEttinger
@RayzJulia