If I lean on Ernie my back will hurt less.
Elmo will feel appreciated if I give him a flower.

If I lean on Ernie my back will hurt less.
Elmo will feel appreciated if I give him a flower

If I lean on Ernie my back will hurt less

om nom nom!
Do pre-trained LMs *already* capture commonsense knowledge?
To fine-tune or not to fine-tune, that is the question
To fine-tune or not to fine-tune, that is the question.
Knowledge-base Completion

Converting KB relations to natural language templates and using LMs to query / score

LMs:
Templates:
KBs:
Conclusion:
Knowledge-base Completion

Converting KB relations to natural language templates and using LMs to query / score

- **Petroni et al. (2019):**
  - **LMs:** ELMo / BERT
  - **Templates:** Hand-crafted templates
  - **KBs:** ConceptNet and Wikidata
  - **Conclusion:** BERT performs well but all models perform poorly on many-to-many relations

![Diagram](image)
Knowledge-base Completion

Converting KB relations to natural language templates and using LMs to query / score

- **Petroni et al. (2019):**
  - LMs: ELMo / BERT
  - Templates: Hand-crafted templates
  - KBs: ConceptNet and Wikidata
  - Conclusion: BERT performs well but all models perform poorly on many-to-many relations

- **Feldman et al. (2019):**
  - LMs: BERT
  - Templates: Hand-crafted templates scored by GPT2
  - KBs: ConceptNet, mining from Wikipedia
  - Conclusion: Performs worse than supervised methods on ConceptNet but is more likely to generalize to different domains

---

**Table 1:** Example of generating candidate sentences. Several enumerated sentences for the triple (musician, CapableOf, play musical instrument). The sentence with the highest log-likelihood according to a pretrained language model is selected.

<table>
<thead>
<tr>
<th>Candidate Sentence $S_i$</th>
<th>log $p(S_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>“musician can playing musical instrument”</td>
<td>$-5.7$</td>
</tr>
<tr>
<td>“musician can be play musical instrument”</td>
<td>$-4.9$</td>
</tr>
<tr>
<td>“musician often play musical instrument”</td>
<td>$-5.5$</td>
</tr>
<tr>
<td>“a musician can play a musical instrument”</td>
<td>$-2.9$</td>
</tr>
</tbody>
</table>
Properties of Concepts (Weir et al., 2020)

1) Do pre-trained LMs correctly distinguish concepts associated with a given set of assumed properties?
Properties of Concepts (Weir et al., 2020)

1) Do pre-trained LMs correctly distinguish concepts associated with a given set of assumed properties?

A ___ has fur.
Properties of Concepts (Weir et al., 2020)

1) Do pre-trained LMs correctly distinguish concepts associated with a given set of assumed properties?

A ___ has fur.
1) Do pre-trained LMs correctly distinguish concepts associated with a given set of assumed properties?

A ___ has fur.
A ___ has fur, is big, and has claws.
Properties of Concepts (Weir et al., 2020)

1) Do pre-trained LMs correctly distinguish concepts associated with a given set of assumed properties?

A ___ has fur.
A ___ has fur, is big, and has claws.
A ___ has fur, is big, and has claws, has teeth, is an animal, ...
Properties of Concepts (Weir et al., 2020)

1) Do pre-trained LMs correctly distinguish concepts associated with a given set of assumed properties?

- Good performance, RoBERTa > BERT
Properties of Concepts (Weir et al., 2020)

1) Do pre-trained LMs correctly distinguish concepts associated with a given set of assumed properties?

- Good performance, RoBERTa > BERT

- Perceptual (e.g. visual) < non-perceptual (e.g. encyclopaedic or functional) - can’t be learned from texts alone
1) Do pre-trained LMs correctly distinguish concepts associated with a given set of assumed properties?

- Good performance, RoBERTa > BERT
- Perceptual (e.g. visual) < non-perceptual (e.g. encyclopaedic or functional) - can’t be learned from texts alone
- Highly-ranked incorrect answers typically apply to a subset of properties
2) Can pre-trained LMs be used to list the properties associated with given concepts?

<table>
<thead>
<tr>
<th>Context</th>
<th>Human Response</th>
<th>RoBERTa-L Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Everyone knows that a bear has</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fur</td>
<td>27</td>
<td>teeth</td>
</tr>
<tr>
<td>claws</td>
<td>15</td>
<td>claws</td>
</tr>
<tr>
<td>teeth</td>
<td>11</td>
<td>eyes</td>
</tr>
<tr>
<td>cubs</td>
<td>7</td>
<td>ears</td>
</tr>
<tr>
<td>paws</td>
<td>7</td>
<td>horns</td>
</tr>
</tbody>
</table>
Properties of Concepts (Weir et al., 2020)

2) Can pre-trained LMs be used to list the properties associated with given concepts?

Low correlation with human elicited properties, but coherent and mostly “verifiable by humans”.

<table>
<thead>
<tr>
<th>Context</th>
<th>Human Response</th>
<th>RoBERTa-L Response</th>
<th>PLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Everyone knows that a</td>
<td>fur 27</td>
<td>teeth .36</td>
<td></td>
</tr>
<tr>
<td>bear has ____</td>
<td>claws 15</td>
<td>claws .18</td>
<td></td>
</tr>
<tr>
<td></td>
<td>teeth 11</td>
<td>eyes .05</td>
<td></td>
</tr>
<tr>
<td></td>
<td>cubs 7</td>
<td>ears .03</td>
<td></td>
</tr>
<tr>
<td></td>
<td>paws 7</td>
<td>horns .02</td>
<td></td>
</tr>
</tbody>
</table>
Can we trust knowledge from LMs?
How well do LMs handle mutual exclusivity?*

Sentence:
The color of the dove who was sitting on the bench was [MASK].

Mask 1 Predictions:
- 15.0% red
- 9.8% blue
- 7.0% different
- 5.7% yellow
- 5.3% purple

https://demo.allennlp.org/masked-lm
LMs also generate fictitious facts!
LMs also generate fictitious facts!

Barack’s Wife Hillary: Using Knowledge Graphs for Fact-Aware Language Modeling

Robert L. Logan IV* Nelson F. Liu† Matthew E. Peters§
Matt Gardner§ Sameer Singh*

* University of California, Irvine, CA, USA
† University of Washington, Seattle, WA, USA
§ Allen Institute for Artificial Intelligence, Seattle, WA, USA

{rlogan, sameer}@uci.edu, {mattg, matthewp}@allenai.org, nfliu@cs.washington.edu
LMs also generate fictitious facts!

Barack’s Wife Hillary: Using Knowledge Graphs for Fact-Aware Language Modeling

Robert L. Logan IV* Nelson F. Liu† Matthew E. Peters§
Matt Gardner§ Sameer Singh*

* University of California, Irvine, CA, USA
† University of Washington, Seattle, WA, USA
§ Allen Institute for Artificial Intelligence, Seattle, WA, USA

{rlogan, sameer}@uci.edu, {mattg, matthewp}@allenai.org, nfliu@cs.washington.edu

Negated and Mismarked Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly

Nora Kassner, Hinrich Schütze
Center for Information and Language Processing (CIS)
LMU Munich, Germany
kassner@cis.lmu.de
Zero-shot LM-based Models for commonsense tasks
Zero-shot setup
Zero-shot setup

\[ P_{LM}(\text{The answer is } \text{answer\_choice\_1}) \]
\[ P_{LM}(\text{The answer is } \text{answer\_choice\_2}) \]

\[ \ldots \]

\[ P_{LM}(\text{The answer is } \text{answer\_choice\_k}) \]

Language Model
Zero-shot setup

\[
\begin{align*}
&\text{Language Model} \\
&P_{\text{LM}}(\text{The answer is } \text{answer choice}_1) \\
&P_{\text{LM}}(\text{The answer is } \text{answer choice}_2) \\
&\quad \ldots \\
&P_{\text{LM}}(\text{The answer is } \text{answer choice}_k) \\
\end{align*}
\]

\[
\begin{align*}
&\text{Masked Language Model} \\
&P_{\text{LM}}(\text{answer choice}_1 \mid \text{The answer is } [\text{MASK}]) \\
&P_{\text{LM}}(\text{answer choice}_2 \mid \text{The answer is } [\text{MASK}]) \\
&\quad \ldots \\
&P_{\text{LM}}(\text{answer choice}_k \mid \text{The answer is } [\text{MASK}]) \\
\end{align*}
\]
Can we use LMs to generate required, missing or implicit knowledge for multiple choice commonsense question answering tasks?
What do professors primarily do? Teach courses. The main function of a professor’s teaching career is to teach students how they can improve their knowledge.

What do professors primarily do? Wear wrinkled tweed jackets. The main function of a professor’s teaching career and is to provide instruction in the subjects they teach.
Generating Clarifications

**Question Generation**

*What do professors primarily do?*
Generating Clarifications

Question Generation

What do professors primarily do?

What is the main function of

DistilGPT2
Generating Clarifications

Question Generation

What do professors primarily do?

What is the main function of

Clarification Generation

What do professors primarily do?

What is the main function of a professor’s teaching career?

The main function of a professor’s teaching career is

DistilGPT2

a professor’s teaching career?
The main function of a professor’s teaching career is to teach students how they can improve their knowledge.
Knowledge-informed Model

Generating clarifications from ConceptNet, Google Ngrams and COMET

Taylor was doing her job so she put the money in the drawer.

What will Taylor want to do next?
Taylor was doing her job so she put the money in the drawer.

Job is a type of work. You would work because you want money.

What will Taylor want to do next?
Knowledge-informed Model

Generating clarifications from ConceptNet, Google Ngrams and COMET

Taylor was doing her job so she put the money in the drawer.

What will Taylor want to do next?

Job is a type of work. You would work because you want money.

Job to earn money.
Taylor was doing her job so she put the money in the drawer.

Job is a type of work. You would work because you want money.

Job to earn money.

As a result, Taylor wants to keep the money in the drawer.

What will Taylor want to do next?

Job to earn money.

To keep the money in the drawer.
Unsupervised Commonsense Question Answering with Self-Talk

- Generating knowledge with LMs improve upon the baseline and performs similarly to knowledge-informed models.
Unsupervised Commonsense Question Answering with Self–Talk

- Generating knowledge with LMs improve upon the baseline and performs similarly to knowledge-informed models.

- Generated clarifications don’t align with what humans consider helpful.
### Unsupervised Commonsense Question Answering with Self-Talk

- Generating knowledge with LMs improve upon the baseline and performs similarly to knowledge-informed models.

- Generated clarifications don’t align with what humans consider helpful.

<table>
<thead>
<tr>
<th>WinoGrande</th>
<th>SocialIQa</th>
<th>MC-TACO</th>
<th>PIQA</th>
<th>CommonSenseQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevant</td>
<td>Relevant</td>
<td>Relevant</td>
<td>Relevant</td>
<td>Relevant</td>
</tr>
<tr>
<td>Correct</td>
<td>Correct</td>
<td>Correct</td>
<td>Correct</td>
<td>Correct</td>
</tr>
<tr>
<td>Helpful</td>
<td>Helpful</td>
<td>Helpful</td>
<td>Helpful</td>
<td>Helpful</td>
</tr>
</tbody>
</table>

- COMET: [Bar chart showing performance metrics for COMET across different datasets]
- ConceptNet: [Bar chart showing performance metrics for ConceptNet across different datasets]
- Distil-GPT2: [Bar chart showing performance metrics for Distil-GPT2 across different datasets]
- GPT2: [Bar chart showing performance metrics for GPT2 across different datasets]
- GPT2-M: [Bar chart showing performance metrics for GPT2-M across different datasets]
- GPT2-XL: [Bar chart showing performance metrics for GPT2-XL across different datasets]
- GPT2-L: [Bar chart showing performance metrics for GPT2-L across different datasets]
- GPT: [Bar chart showing performance metrics for GPT across different datasets]
- XLNet-base: [Bar chart showing performance metrics for XLNet-base across different datasets]
- XLNet-L: [Bar chart showing performance metrics for XLNet-L across different datasets]
To fine-tune or not to fine-tune, that is the question
LMs provide a good basis for commonsense task models

**MC-TACO**

MC-TACO is a dataset of 13k question-answer pairs that require temporal commonsense comprehension... [More]

**Human Performance**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Submission</th>
<th>Created</th>
<th>Exact Match</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>T5 - 3B fine-tuned + number n... Zakaria Kaddari, Youssef Mell...</td>
<td>03/24/2020</td>
<td>0.5908</td>
<td>0.7946</td>
</tr>
<tr>
<td>2</td>
<td>T5 - 3B baseline Zakaria Kaddari, Youssef Mell...</td>
<td>03/15/2020</td>
<td>0.5758</td>
<td>0.7845</td>
</tr>
</tbody>
</table>

Exact Match: 0.7580
LMs provide a good basis for commonsense task models

aNLI
Abductive Natural Language Inference (aNLI) is a new commonsense benchmark dataset designed to test... (More)

Human Performance
Accuracy: 0.9290

<table>
<thead>
<tr>
<th>Rank</th>
<th>Submission</th>
<th>Created</th>
<th>Run Time</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>L2R2 (RoBERTa Large + Likeli...)</td>
<td>01/27/2020</td>
<td>33 minutes</td>
<td>0.8681</td>
</tr>
<tr>
<td>2</td>
<td>egel</td>
<td>01/20/2020</td>
<td>29 minutes</td>
<td>0.8595</td>
</tr>
</tbody>
</table>
LMs provide a good basis for commonsense task models

Physical IQA

We introduce Physical IQA: Physical Interaction QA, a new commonsense QA benchmark for naive... (More)
LMs provide a good basis for commonsense task models

WinoGrande

WinoGrande is a new collection of 44k problems, inspired by Winograd Schema Challenge (Levesque... (More)

Human Performance

<table>
<thead>
<tr>
<th>Rank</th>
<th>Submission</th>
<th>Created</th>
<th>AUC</th>
<th>Acc (XS)</th>
<th>Acc (S)</th>
<th>Acc (M)</th>
<th>Acc (L)</th>
<th>Acc (XL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TTTTT</td>
<td>03/13/2020</td>
<td>0.7673</td>
<td>0.6825</td>
<td>0.7051</td>
<td>0.7759</td>
<td>0.8240</td>
<td>0.8461</td>
</tr>
<tr>
<td>2</td>
<td>Roberta-large + G-DAug-Combo</td>
<td>02/23/2020</td>
<td>0.7146</td>
<td>0.6106</td>
<td>0.6712</td>
<td>0.7119</td>
<td>0.7736</td>
<td>0.7923</td>
</tr>
<tr>
<td>3</td>
<td>Roberta-large + G-DAug-Div</td>
<td>02/16/2020</td>
<td>0.7118</td>
<td>0.6163</td>
<td>0.6667</td>
<td>0.7046</td>
<td>0.7714</td>
<td>0.7929</td>
</tr>
</tbody>
</table>
LMs provide a good basis for commonsense task models

...but they need a “push in the right direction” (fine tuning)
Can good performance be attributed to knowledge in LMs or to training a large model on a large dataset?
HellaSwag (Zellers et al., 2019)
HellaSwag (Zellers et al., 2019)

- LMs mostly pick up *lexical cues*
- No model actually solves commonsense reasoning to date.

A woman is outside with a bucket and a dog. The dog is running around trying to avoid a bath. She...

A. rinses the bucket off with soap and blow dry the dog’s head.
B. uses a hose to keep it from getting soapy.
C. gets the dog wet, then it runs away again.
D. gets into a bath tub with the dog.
HellaSwag (Zellers et al., 2019)

- LMs mostly pick up *lexical cues*
- No model actually solves commonsense reasoning to date.

If no algorithmic advance is made, it would take **100k GPU hours** to reach human performance on HellaSWAG!
PIQA (Bisk et al., 2020)

LMs lack an understanding of some of the most basic physical properties of the world.

To separate egg whites from the yolk using a water bottle, you should...

a. **Squeeze** the water bottle and press it against the yolk. **Release**, which creates suction and lifts the yolk.

b. **Place** the water bottle and press it against the yolk. **Keep pushing**, which creates suction and lifts the yolk.
Can you teach LMs commonsense?
Do Neural Language Representations Learn Physical Commonsense?

Forbes et al. (2019): Fine-tune BERT to predict object properties ("uses electricity"), affordances ("plug in"), and the inferences between them (e.g. plug-in(x) ⇒ x uses electricity).
Forbes et al. (2019): Fine-tune BERT to predict object properties ("uses electricity"), affordances ("plug in"), and the inferences between them (e.g. plug-in(x)⇒x uses electricity).

**Best performance**: functional properties (e.g. "uses electricity") given affordances.

**Reasonable performance**: encyclopedic (is an animal) and commonsense properties (comes in pairs).

**Worst performance**: perceptual properties (smooth) which are often not expressed by affordances.
Can you teach LMs symbolic reasoning?

Talmor et al. (2019): oLMpics - testing BERT and RoBERTa on a set of symbolic reasoning tasks:
Can you teach LMs symbolic reasoning?

Talmor et al. (2019): oLMpics - testing BERT and RoBERTa on a set of symbolic reasoning tasks:

**Always-Never:** A chicken [MASK] has horns.  
A. never  
B. rarely  
C. sometimes  
D. often  
E. always
Can you teach LMs symbolic reasoning?

Talmor et al. (2019): oLMpics - testing BERT and RoBERTa on a set of symbolic reasoning tasks:

**Always-Never:** A chicken [MASK] has horns.  
A. never  
B. rarely  
C. sometimes  
D. often  
E. always

![Image of crossed-out options](Image)
Can you teach LMs symbolic reasoning?

Talmor et al. (2019): oLMpics - testing BERT and RoBERTa on a set of symbolic reasoning tasks:

**Always-Never:** A chicken [MASK] has horns.  
A. never  
B. rarely  
C. sometimes  
D. often  
E. always

Reporting bias: LMs are trained on texts describing things that **do** happen!
Can you teach LMs symbolic reasoning?

Talmor et al. (2019): oLMpics - testing BERT and RoBERTa on a set of symbolic reasoning tasks:

Age Comparison:
A 21 year old person age is [MASK] than a 35 year old person.  
A. younger  B. older
Can you teach LMs symbolic reasoning?

Talmor et al. (2019): oLMpics - testing BERT and RoBERTa on a set of symbolic reasoning tasks:

Age Comparison:
A 21 year old person age is [MASK] than a 35 year old person.  
A. younger  B. older
Can you teach LMs symbolic reasoning?

Talmor et al. (2019): oLMpics - testing BERT and RoBERTa on a set of symbolic reasoning tasks:

**Age Comparison:**
A 21 year old person age is [MASK] than a 35 year old person.  
A. younger  
B. older

RoBERTa also performs well in a zero-shot setup:
Can you teach LMs symbolic reasoning?

Talmor et al. (2019): oLMpics - testing BERT and RoBERTa on a set of symbolic reasoning tasks:

**Negation:** It was [MASK] hot, it was really cold  
A. really  
B. not
Can you teach LMs symbolic reasoning?

<table>
<thead>
<tr>
<th></th>
<th>RoBERTa-L</th>
<th>BERT-WWM</th>
<th>BERT-L</th>
<th>RoBERTa-B</th>
<th>BERT-B</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALWAYS-NEVER</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGE COMPARISON</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OBJECTS COMPARISON</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANTONYM NEGATION</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PROPERTY CONJUNCTION</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TAXONOMY CONJUNCTION</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>ENCYC. COMPOSITION</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>MULTI-HOP COMPARISON</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>
Can you teach LMs symbolic reasoning?

<table>
<thead>
<tr>
<th></th>
<th>RoBERTa-L</th>
<th>BERT-WWM</th>
<th>BERT-L</th>
<th>RoBERTa-B</th>
<th>BERT-B</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ALWAYS-NEVER</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>AGE COMPARISON</strong></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>OBJECTS COMPARISON</strong></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ANTONYM NEGATION</strong></td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>PROPERTY CONJUNCTION</strong></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>TAXONOMY CONJUNCTION</strong></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ENCYC. COMPOSITION</strong></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>MULTI-HOP COMPARISON</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

RoBERTa > BERT
Can you teach LMs symbolic reasoning?

RoBERTa > BERT

<table>
<thead>
<tr>
<th></th>
<th>RoBERTa-L</th>
<th>BERT-WWM</th>
<th>BERT-L</th>
<th>RoBERTa-B</th>
<th>BERT-B</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALWAYS-NEVER</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGE COMPARISON</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OBJECTS COMPARISON</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANTONYM NEGATION</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PROPERTY CONJECTION</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>TAXONOMY CONJUNCTION</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>ENCYC. COMPOSITION</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>MULTI-HOP COMPARISON</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Worse performance on compositionality tasks
Can you teach LMs symbolic reasoning?

RoBERTa > BERT

Worse performance on compositionality tasks

LMs are context-dependent and small changes to the input hurts their performance.
Summary

- Pre-trained language models some commonsense knowledge - but it is far from an exhaustive source.

**Insufficient coverage** (reporting bias; Gordon and Van Durme, 2013).

- Factual world knowledge from pre-trained LMs
Summary

- Pre-trained language models some commonsense knowledge - but it is far from an exhaustive source.

- Use with caution! LMs also generate false facts.

Insufficient coverage (reporting bias; Gordon and Van Durme, 2013).

Insufficient precision from pre-trained LMs.

Factual world knowledge from pre-trained LMs.
Pre-trained language models some commonsense knowledge - but it is far from an exhaustive source.

Use with caution! LMs also generate false facts.

Thank you! Questions?

vereds@allenai.org
References + Additional Reading


