Commonsense benchmarks

Or how to measure that your model is actually doing some commonsense reasoning
How do you know that a model is doing commonsense reasoning?
How do you know that a model is doing commonsense reasoning?

**Unsupervised:**
- Observe behavior,
- Probe representations,
- etc.
How do you know that a model is doing commonsense reasoning?

**Unsupervised:**
- Observe behavior,
- Probe representations,
- etc.

**Benchmarks:**
knowledge-specific tests (w/ or w/o training data)
How do you know that a model is doing commonsense reasoning?

**Unsupervised:**
- Observe behavior,
- Probe representations,
- etc.

**Benchmarks:**
knowledge-specific tests (w/ or w/o training data)

**QA format:** easy to evaluate (e.g., accuracy)
Step 1: Determine type of reasoning

https://leaderboard.allenai.org/
Step 1: Determine type of reasoning

- Abductive reasoning
- Physical IQa
- Social IQa
- VCR
- HELLA SWAG

https://leaderboard.allenai.org/
Step 1: Determine type of reasoning

Abductive reasoning

Visual commonsense reasoning

Abductive NLI  Physical IQa  Social IQa  VCR  HELLA SWAG

https://leaderboard.allenai.org/
Reasoning about Social Situations
Reasoning about Social Situations

Alex spilt food all over the floor and it made a huge mess.

What will Alex want to do next?
Alex spilt food all over the floor and it made a huge mess.

What will Alex want to do next?

- run around in the mess
- mop up the mess
Alex spilt food all over the floor and it made a huge mess.

What will Alex want to do next?

- run around in the mess
- mop up the mess

less likely

more likely
Knowledge tested in **SOCIAL IQA: ATOMIC**

**causes**
- drink too much
- fall over
- no intent
- X needed to
- X wanted to

**stative**
- clumsy
- careless
- X is seen as
- has effect on X

**effects**
- embarrassed
- upset
- X will feel
- X will want
- clean it up
- get a broom
- slip on the spill
- gets dirty

**PersonX spills ___ all over the floor**
### Step 2: Choosing a benchmark size

<table>
<thead>
<tr>
<th></th>
<th><strong>Small scale</strong></th>
<th><strong>Large scale</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Creation</strong></td>
<td>Expert-curated</td>
<td>Crowdsourced/automatic</td>
</tr>
<tr>
<td><strong>Coverage</strong></td>
<td>Limited coverage</td>
<td>Large coverage</td>
</tr>
<tr>
<td><strong>Training</strong></td>
<td>Dev/test only</td>
<td>Training/dev/test</td>
</tr>
<tr>
<td><strong>Budget</strong></td>
<td>Expert time costs</td>
<td>Crowdsourcing costs</td>
</tr>
</tbody>
</table>
## Step 2: Choosing a benchmark size

<table>
<thead>
<tr>
<th></th>
<th>Small scale</th>
<th>Large scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creation</td>
<td>Expert-curated</td>
<td>Crowdsourced/automatic</td>
</tr>
<tr>
<td>Coverage</td>
<td>Limited coverage</td>
<td>Large coverage</td>
</tr>
<tr>
<td>Training</td>
<td>Dev/test only</td>
<td>Training/dev/test</td>
</tr>
<tr>
<td>Budget</td>
<td>Expert time costs</td>
<td>Crowdsourcing costs</td>
</tr>
</tbody>
</table>

Winograd Schema Challenge (WSC), Choice of Plausible Alternatives (COPA)
Small commonsense benchmarks

Winograd Schema Challenge (WSC)
273 examples

Choice of Plausible Alternatives (COPA)
500 dev, 500 test

The city councilmen refused the demonstrators a permit because they advocated violence. Who is “they”?

(a) The city councilmen
(b) The demonstrators

The city councilmen refused the demonstrators a permit because they feared violence. Who is “they”?

(a) The city councilmen
(b) The demonstrators
Small commonsense benchmarks

Winograd Schema Challenge (WSC)
273 examples

Choice of Plausible Alternatives (COPA)
500 dev, 500 test

I hung up the phone.
What was the cause of this?

(a) The caller said goodbye to me.
(b) The caller identified himself to me.

The toddler became cranky.
What happened as a result?

(a) Her mother put her down for a nap.
(b) Her mother fixed her hair into pigtails.
### Step 2: Choosing a QA benchmark size

<table>
<thead>
<tr>
<th></th>
<th>Small scale</th>
<th>Large scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creation</td>
<td>Expert-curated</td>
<td>Crowdsourced/automatic</td>
</tr>
<tr>
<td>Coverage</td>
<td>Limited coverage</td>
<td>Large coverage</td>
</tr>
<tr>
<td>Training</td>
<td>Dev/test only</td>
<td>Training/dev/test</td>
</tr>
<tr>
<td>Budget</td>
<td>Expert time costs</td>
<td>Crowdsourcing costs</td>
</tr>
</tbody>
</table>

**Challenge**: do to collect positive/negative answers?
Challenge of collecting unlikely answers
Challenge of collecting unlikely answers

**Goal:** negative answers have to be *plausible but unlikely*
Challenge of collecting unlikely answers

**Goal**: negative answers have to be *plausible but unlikely*

- Automatic matching?
  - Random negative sampling won’t work, too topically different
  - “smart” negative sampling isn’t effective either
Challenge of collecting unlikely answers

**Goal**: negative answers have to be *plausible but unlikely*

- Automatic matching?
  - Random negative sampling won’t work, too topically different
  - “smart” negative sampling isn’t effective either
- Need better solution... maybe we can ask crowd workers?
Alex spilt food all over the floor and it made a huge mess.

What will Alex want to do next?
Alex spilt food all over the floor and it made a huge mess.

What will Alex want to do next?
Collecting answers from crowdworkers

<table>
<thead>
<tr>
<th>Context and Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alex spilt food all over the floor and it made a huge mess.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>WHAT HAPPENS NEXT</th>
</tr>
</thead>
<tbody>
<tr>
<td>What will Alex want to do next?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Free Text Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Handwritten ✔ and ✗ Answers</td>
</tr>
<tr>
<td>✔ mop up</td>
</tr>
<tr>
<td>✔ give up and order take out</td>
</tr>
<tr>
<td>✗ leave the mess</td>
</tr>
<tr>
<td>✗ run around in the mess</td>
</tr>
</tbody>
</table>
Collecting answers from crowdworkers

Context and Question

Alex spilt food all over the floor and it made a huge mess. What will Alex want to do next?

WHAT HAPPENS NEXT

Free Text Response

Handwritten ✔ and ✘ Answers

✔ mop up
✔ give up and order take out
✘ leave the mess
✘ run around in the mess

Problem: handwritten unlikely answers are too easy to detect
Problem: annotation artifacts
Problem: annotation artifacts

- Models can exploit artifacts in handwritten incorrect answers
  - Exaggerations, off-topic, overly emotional, etc.
Problem: annotation artifacts

- Models can exploit artifacts in handwritten incorrect answers
  - Exaggerations, off-topic, overly emotional, etc.
- Seemingly “super-human” performance by large pretrained LMs (BERT, GPT, etc.)
Problem: annotation artifacts

• Models can exploit artifacts in handwritten incorrect answers
  • Exaggerations, off-topic, overly emotional, etc.
• Seemingly “super-human” performance by large pretrained LMs (BERT, GPT, etc.)
How to make unlikely answers robust to annotation artifacts?
How to make unlikely answers robust to annotation artifacts?

**SOCIAL IQA, COMMONSENSEQA:** Modified answer collection
How to make unlikely answers robust to annotation artifacts?

**SOCIAL IQA, COMMONSENSEQA**: Modified answer collection

**HellaSwag & AF-lite**: Adversarial filtering of artifacts
Alex spilt food all over the floor and it made a huge mess.

What will Alex want to do next?

- ✔ mop up
- ✔ give up and order take out
- ✗ have slippery hands
- ✗ get ready to eat
Question-Switching Answers (SOCIAL IQA)

Original Question
Alex spilt food all over the floor and it made a huge mess.

WHAT HAPPENS NEXT
What will Alex want to do next?
✔ mop up
✔ give up and order take out
❌

Question-Switching Answer
WHAT HAPPENED BEFORE
What did Alex need to do before this?

Original Question
Alex spilt food all over the floor and it made a huge mess.

WHAT HAPPENS NEXT
What will Alex want to do next?
✔ mop up
✔ give up and order take out
❌

Question-Switching Answer
WHAT HAPPENED BEFORE
What did Alex need to do before this?

Original Question
Alex spilt food all over the floor and it made a huge mess.

WHAT HAPPENS NEXT
What will Alex want to do next?
✔ mop up
✔ give up and order take out
❌

Question-Switching Answer
WHAT HAPPENED BEFORE
What did Alex need to do before this?

Original Question
Alex spilt food all over the floor and it made a huge mess.

WHAT HAPPENS NEXT
What will Alex want to do next?
✔ mop up
✔ give up and order take out
❌

Question-Switching Answer
WHAT HAPPENED BEFORE
What did Alex need to do before this?

Original Question
Alex spilt food all over the floor and it made a huge mess.

WHAT HAPPENS NEXT
What will Alex want to do next?
✔ mop up
✔ give up and order take out
❌

Question-Switching Answer
WHAT HAPPENED BEFORE
What did Alex need to do before this?

Original Question
Alex spilt food all over the floor and it made a huge mess.

WHAT HAPPENS NEXT
What will Alex want to do next?
✔ mop up
✔ give up and order take out
❌

Question-Switching Answer
WHAT HAPPENED BEFORE
What did Alex need to do before this?
Question-Switching Answers (SOCIAL IQA)

Original Question

Alex spilt food all over the floor and it made a huge mess.

**WHAT HAPPENS NEXT**
What will Alex want to do next?

- ✔ mop up
- ✔ give up and order take out
- ✗
- ✗

Question-Switching Answer

**WHAT HAPPENED BEFORE**
What did Alex need to do before this?

- ✔ have slippery hands
- ✔ get ready to eat

- ✗
- ✗
Alex spilt food all over the floor and it made a huge mess. What will Alex want to do next?

✔ mop up
✔ give up and order take out
❌ have slippery hands
❌ get ready to eat

What happened before this?

✔ have slippery hands
✔ get ready to eat
Comparing incorrect/correct answers’ styles

More stylistically different from correct

More stylistically similar
Comparing incorrect/correct answers’ styles

Effect Size when comparing to Correct Answers

- More stylistically different from correct
- More stylistically similar

Handwritten Incorrect | Question Switching
---|---
Arousal | Dominance | Valence
0.05 | 0.15 | 0.40
Comparing incorrect/correct answers’ styles

Effect Size when comparing to Correct Answers

Question switching answers are more stylistically similar to correct answers.
COMMONSENSEQA: pivot on knowledge graphs

Filter edges from ConceptNet with rules
COMMONSENSEQA: pivot on knowledge graphs

Filter edges from ConceptNet with rules

Extract subgraphs from ConceptNet

- dust
- attic
- yard
- street
- glass
- bar
- fork
- car
- happy
- laugh
- sad
- fall

Talmor et al. 2019
COMMONSENSEQA: pivot on knowledge graphs

Filter edges from ConceptNet with rule

Extract subgraphs from ConceptNet

Crowdworkers author questions

Dust in house? (attic, yard, street)
Find glass outside? (bar, fork, car)
Makes you happy? (laugh, sad, fall)

Talmor et al. 2019
COMMONSENSEQA: pivot on knowledge graphs

Filter edges from ConceptNet with rules:

Extract subgraphs from ConceptNet:
- dust
- attic
- yard
- street
- glass
- bar
- fork
- car
- happy
- laugh
- sad
- fall

Crowdworkers author questions:
- Dust in house? (attic, yard, street)
- Find glass outside? (bar, fork, car)

Crowdworkers add distractors:
- Dust in house? (attic, yard, street, bed, desert)
- Find glass outside? (bar, fork, car, sand, wine)
- Makes you happy? (laugh, sad, fall, blue, feel)

Talmor et al. 2019
COMMONSENSEQA: pivot on knowledge graphs

Filter edges from ConceptNet with rules

Extract subgraphs from ConceptNet

Crowdworkers author questions
- Dust in house? (attic, yard, street)
- Find glass outside? (bar, fork, car)

Crowdworkers add distractors
- Dust in house? (attic, yard, street, bed, desert)
- Find glass outside? (bar, fork, car, sand, wine)

Crowdworkers filter questions by quality
- Dust in house? (attic, yard, ...) → 1.0
- Find glass outside? (bar, fork, ...) → 0.2
- Makes you happy? (laugh, sad, fall, blue, feel) → 0.8
COMMONSENSEQA: pivot on knowledge graphs

Filter edges from ConceptNet with rules:
- dust
- attic
- yard
- street
- glass
- bar
- fork
- car
- happy
- laugh
- sad
- fall

Collect relevant snippets via search engine:
- Dust in house? (attic, yard, ...)
- Makes you happy? (laugh, sad, ...)

Extract subgraphs from ConceptNet

Crowdworkers author questions:
- Dust in house? (attic, yard, street)
- Find glass outside? (bar, fork, car)

Crowdworkers add distractors:
- Dust in house? (attic, yard, street, bed, desert)
- Find glass outside? (bar, fork, car, sand, wine)

Crowdworkers filter questions by quality:
- Dust in house? (attic, yard, ...) → 1.0
- Find glass outside? (bar, fork, ...) → 0.2
- Makes you happy? (laugh, sad, ...) → 0.8

Talmor et al. 2019
Adversarial Filtering (lite)

**Goal**: remove examples with exploitable artifacts or spurious correlations

- Use pre-trained representations
- Iteratively remove data that’s easiest to predict by a linear classifier (e.g., logistic)
- Robust examples remain

HellaSwag (Zellers et al., 2019)
AF-lite (Le Bras et al., 2019)
Adversarial Filtering (lite)

Goal: remove examples with exploitable artifacts or spurious correlations

• Use pre-trained representations
• Iteratively remove data that’s easiest to predict by a linear classifier (e.g., logistic)
• Robust examples remain

HellaSwag (Zellers et al., 2019)
AF-lite (Le Bras et al., 2019)
Adversarial Filtering (lite)

**Goal**: remove examples with exploitable artifacts or spurious correlations

- Use pre-trained representations
- Iteratively remove data that’s easiest to predict by a linear classifier (e.g., logistic)
- Robust examples remain

HellaSwag (Zellers et al., 2019)
AF-lite (Le Bras et al., 2019)
Adversarial Filtering (lite)

**Goal**: remove examples with exploitable artifacts or spurious correlations

- Use pre-trained representations
- Iteratively remove data that’s easiest to predict by a linear classifier (e.g., logistic)
- Robust examples remain

HellaSwag (Zellers et al., 2019)
AF-lite (Le Bras et al., 2019)
Adversarial Filtering (lite)

**Goal**: remove examples with exploitable artifacts or spurious correlations

- Use pre-trained representations
- Iteratively remove data that’s easiest to predict by a linear classifier (e.g., logistic)
- Robust examples remain

HellaSwag (Zellers et al., 2019)
AF-lite (Le Bras et al., 2019)
Performance of models on the WikiHow portion of HellaSwag (Zellers et al., 2019) with different AF settings and different training models.
Performance of models on the WikiHow portion of HellaSwag (Zellers et al., 2019) with different AF settings and different training models.

Adversarial filtering removes examples with spurious correlations => Task becomes harder.
Model performance on SOCIAL IQA

Accuracy

- **Humans**
- **Bert-large**
- **Bert-base**
- **GPT** (OpenAI)
- **Random**

Accuracy levels: 0.0% to 10.0% to 20.0% to 30.0% to 40.0% to 50.0% to 60.0% to 70.0% to 80.0% to 90.0%
Model performance on SOCIAL IQA

Humans
Bert-large
Bert-base
GPT
Random

Accuracy

>20% gap to improve on
Although Aubrey was older and stronger, they lost to Alex in arm wrestling. How would Alex feel as a result?

- ashamed
- boastful

they need to practice more
Challenging SOCIAL IQA examples for BERT-large

Although Aubrey was older and stronger, they lost to Alex in arm wrestling.

How would Alex feel as a result?
- ashamed
- boastful
  - they need to practice more

What will Remy want to do next?
- lose her credit card
- arrive at a hotel
  - what Remy did before
- get the key from Skylar

Remy gave Skylar, the concierge, her account so that she could check into the hotel.
Although Aubrey was older and stronger, they lost to Alex in arm wrestling.

How would Alex feel as a result?

- ashamed
- boastful
- they need to practice more

What will Remy want to do next?

- lose her credit card
- arrive at a hotel
- get the key from Skylar

Remy gave Skylar, the concierge, her account so that she could check into the hotel.

Need more robust, person-centric reasoning
Challenging SOCIAL IQA examples for BERT-large

Although Aubrey was older and stronger, they lost to Alex in arm wrestling.

How would Alex feel as a result?
- ashamed
- boastful
- they need to practice more

What will Remy want to do next?
- lose her credit card
- arrive at a hotel
- get the key from Skylar

Need more robust, person-centric reasoning

Need better notion of causes vs. effects
Commonsense benchmarks

- Naïve Psychology
- ROC story
- Social IQa
- Physical IQa
- HellaSwag
- SWAG
- WSC
- COPA
- VCR
- CommonsenseQA
- Abductive NLI
- JHU Ordinal Commonsense
- WinoGrande
- MCTaco
- ReCORD
- CosmosQA
- MultiRC
Commonsense benchmarks

Social commonsense
- Naïve Psychology
- ROC story
- Social IQa
- WSC
- COPA
- VCR
- WinoGrande

Physical IQa
- Abductive NLI
- CommonsenseQA

HellaSwag
- SWAG

JHU Ordinal Commonsense
- ReCORD
- CosmosQA

MultiRC

MCTaco
Commonsense benchmarks

Social commonsense
- Naïve Psychology
- ROC story
- Social IQa
- WSC
- COPA
- VCR
- WinoGrande

Physical commonsense
- Physical IQa
- HellaSwag
- SWAG
- JHU Ordinal Commonsense
- Abductive NLI
- CommonsenseQA
- ReCORD
- CosmosQA
- MultiRC
Commonsense benchmarks

Social commonsense
- Naïve Psychology
- ROC story
- Social IQa
- WSC
- COPA
- VCR
- WinoGrande

Physical commonsense
- Physical IQa
- Abductive NLI
- CommonsenseQA
- JHU Ordinal Commonsense
- HellaSwag
- SWAG

Temporal commonsense
- MCTaco

Commonsense reading comprehension
- ReCORD
- CosmosQA
- MultiRC
Commonsense benchmarks

Social commonsense
- Naïve Psychology
- ROC story
- Social IQa
- WinoGrande
- VCR
- WSC

Physical commonsense
- Physical IQa
- HellaSwag
- SWAG

Abductive NLI
- CommonsenseQA
- Abductive NLI

COPA
- Naïve Psychology

WSC

MultiRC

Thanks! Questions?