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

Abductive NLI

Physical IQa

Social IQa

VCR

HELLA SWAG

https://leaderboard.allenai.org/

Step 1: Determine type of reasoning

Abductive reasoning

Abductive NLI

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/

Social
IQa

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?

Social IQa

Reasoning about Social Situations

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

Reasoning about Social Situations

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

What will Alex want to do next?

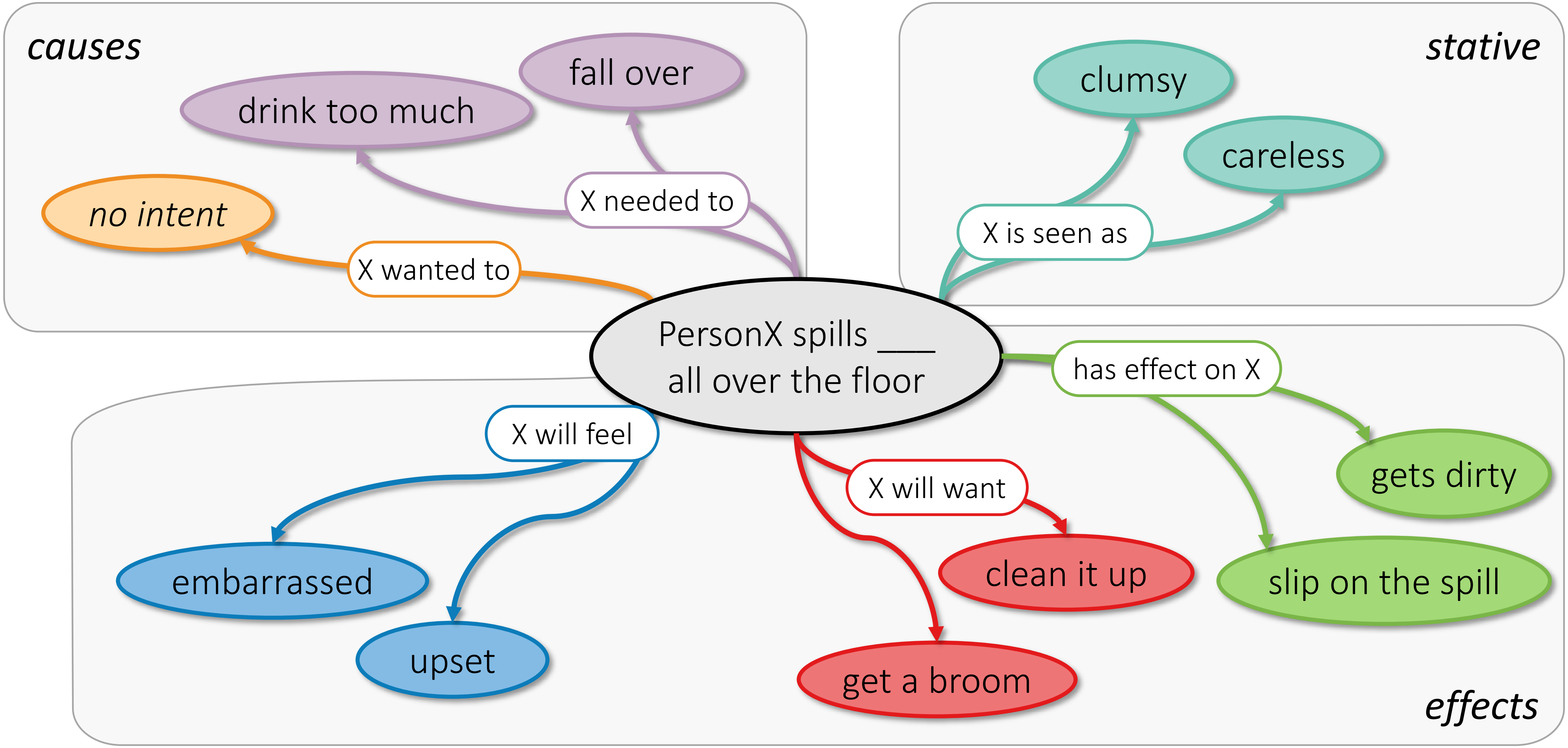
run around in the mess

less likely

mop up the mess

more likely

Knowledge tested in Social IQa: Atomic



Step 2: Choosing a benchmark size

	Small scale	**Large scale**
Creation	Expert-curated	Crowdsourced/automatic
Coverage	Limited coverage	Large coverage
Training	Dev/test only	Training/dev/test
Budget	Expert time costs	Crowdsourcing costs

Step 2: Choosing a benchmark size

	Small scale	Large scale
Creation	Expert-curated	Crowdsourced/automatic
Coverage	Limited coverage	Large coverage
Training	Dev/test only	Training/dev/test
Budget	Expert time costs	Crowdsourcing costs

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

	Small scale	**Large scale**
Creation	Expert-curated	Crowdsourced/automatic
Coverage	Limited coverage	Large coverage
Training	Dev/test only	Training/dev/test
Budget	Expert time costs	Crowdsourcing costs

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?

Collecting answers from crowdworkers

Social IQa

Context and Question

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

WHAT HAPPENS NEXT

What will Alex want to do next?

Social IQa

Collecting answers from crowdworkers

amazon mturk Requester

Context and Question

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

WHAT HAPPENS NEXT

What will Alex want to do next?

Collecting answers from crowdworkers

Context and Question

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

WHAT HAPPENS NEXT

What will Alex want to do next?

Free Text Response

Handwritten ✓ and ✗ Answers

✓ mop up
✓ give up and order take out
✗ leave the mess
✗ run around in the mess

Collecting answers from crowdworkers

Social IQa

Context and Question

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

WHAT HAPPENS NEXT

What will Alex want to do next?

amazon mturk Requester

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.
  - See Schwartz et al. 2017, Gururangan et al. 2018, Zellers et al. 2018, etc.

Problem: annotation artifacts

- Models can exploit artifacts in handwritten incorrect answers
  - Exaggerations, off-topic, overly emotional, etc.
  - See Schwartz et al. 2017, Gururangan et al. 2018, Zellers et al. 2018, 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.
  - See Schwartz et al. 2017, Gururangan et al. 2018, Zellers et al. 2018, 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

Question-Switching Answers (SOCIAL IQA)

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 Answers (SOCIAL IQA)

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?

Question-Switching Answers (Social IQa)

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

Question-Switching Answers (Social IQa)

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
- ✗ have slippery hands
- ✗ get ready to eat

Question-Switching Answer

What happened Before

What did Alex need to do before this?

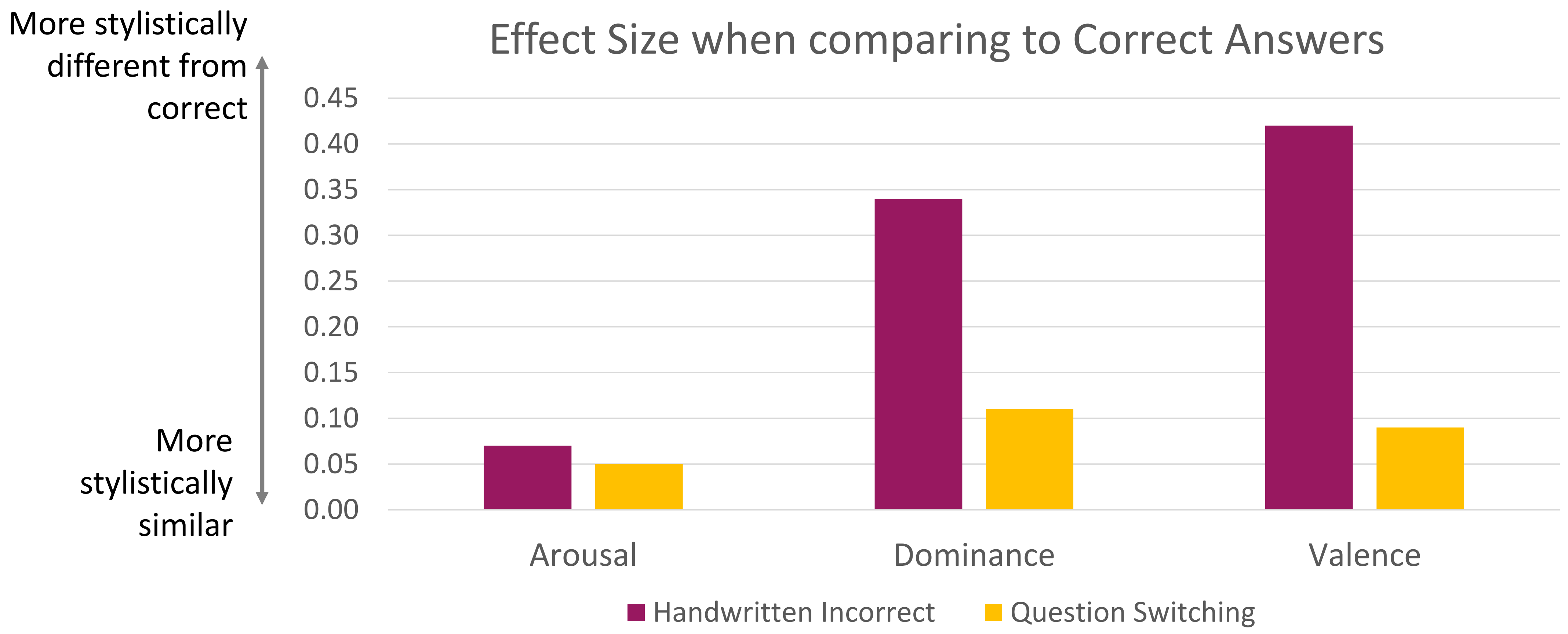
- ✓ have slippery hands
- ✓ get ready to eat

Comparing incorrect/correct answers' styles

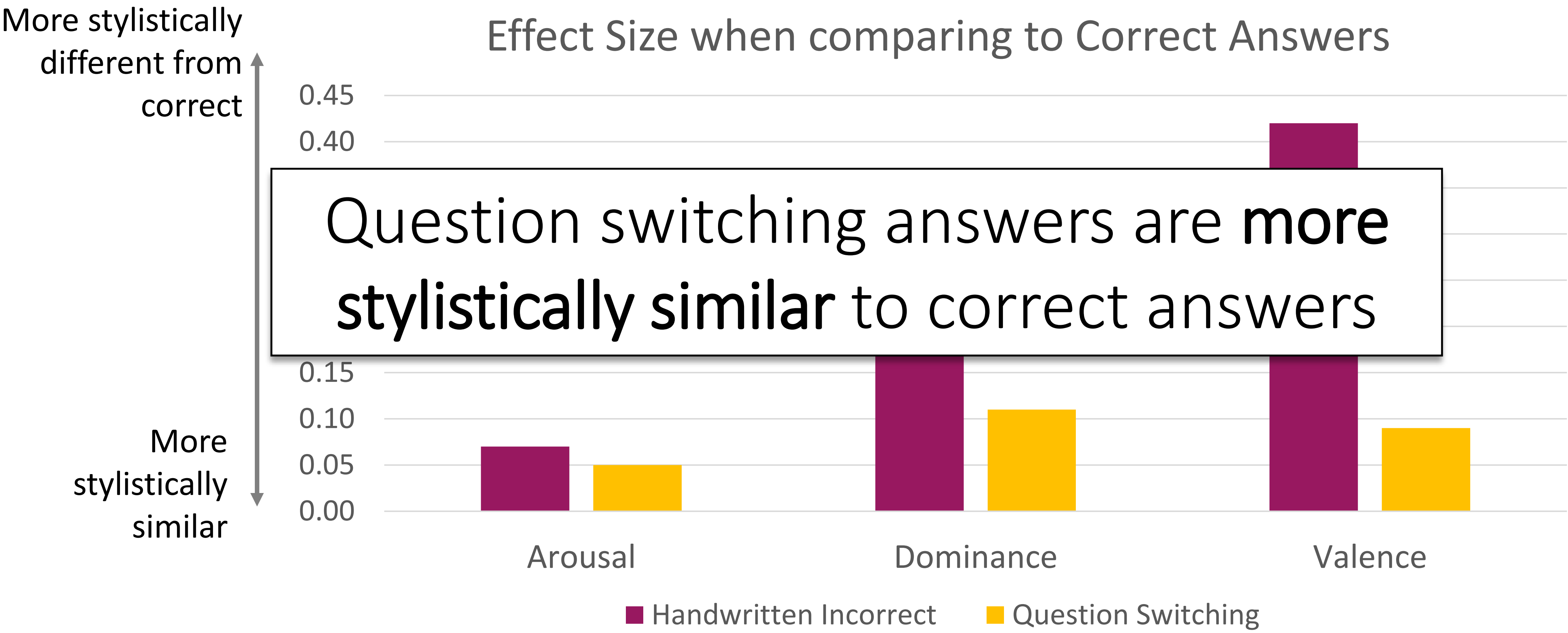
More stylistically different from correct

More stylistically similar

Comparing incorrect/correct answers' styles

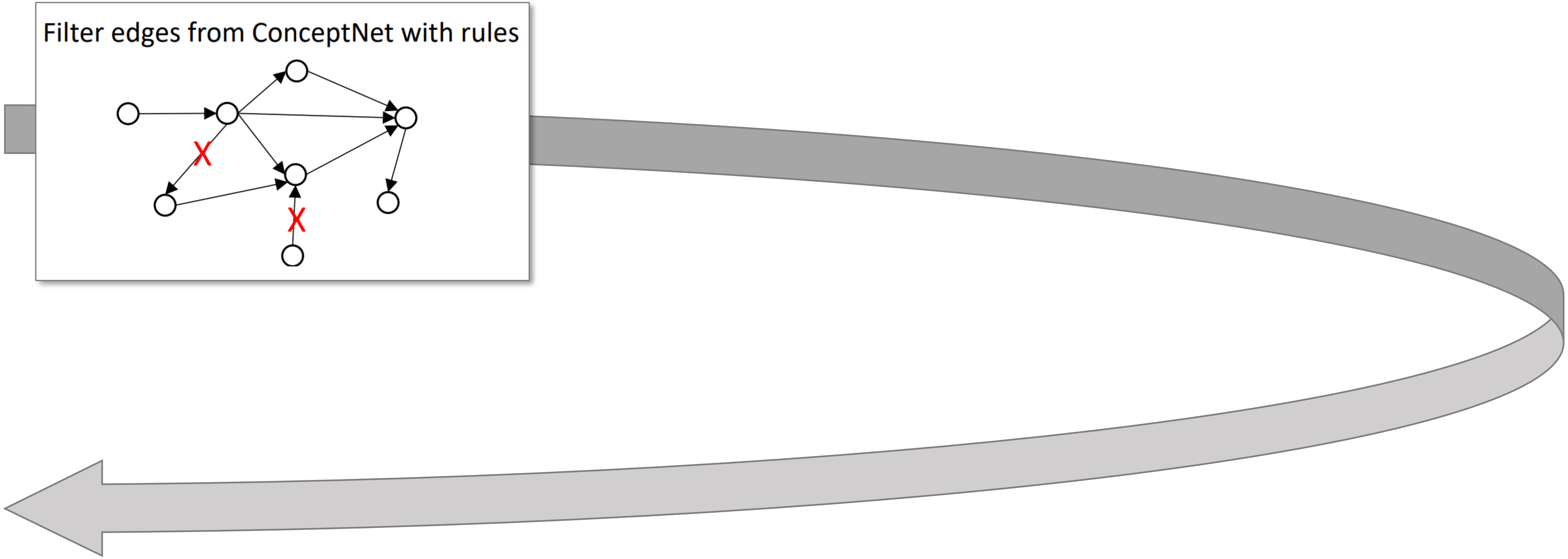


Comparing incorrect/correct answers' styles



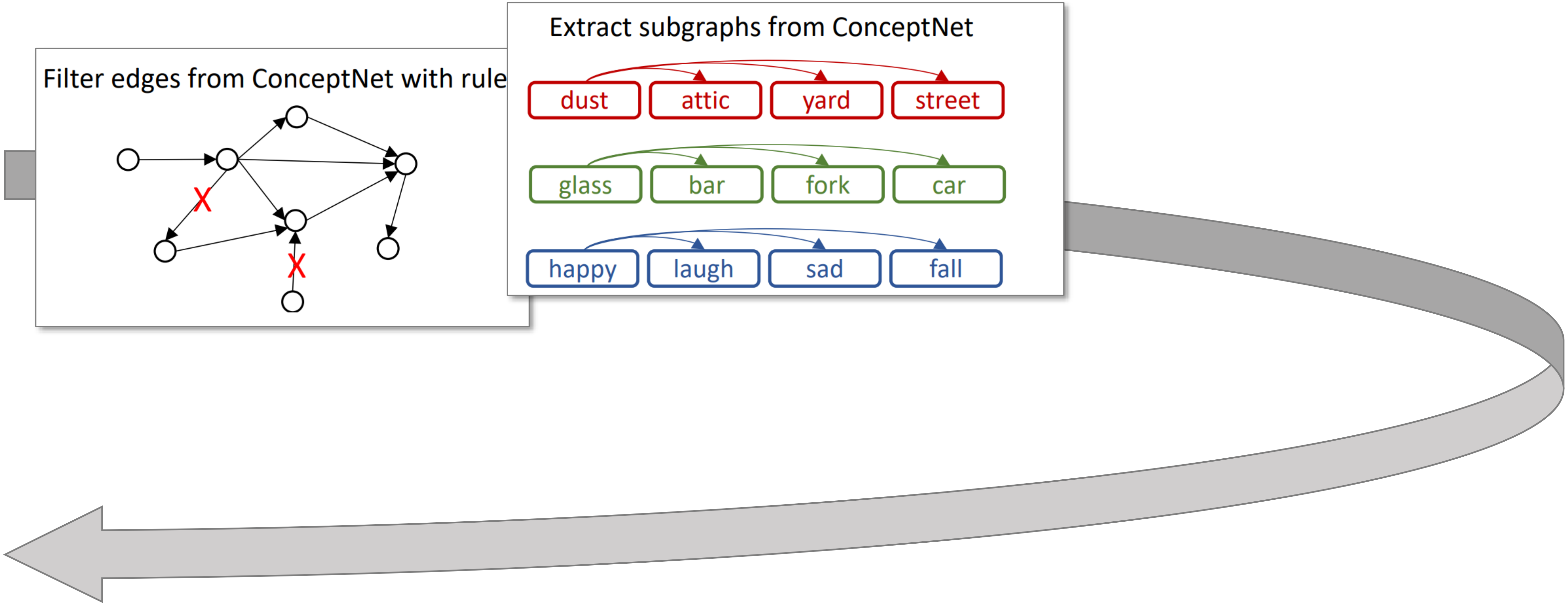
Question switching answers are **more stylistically similar** to correct answers

CommonsenseQA: pivot on knowledge graphs



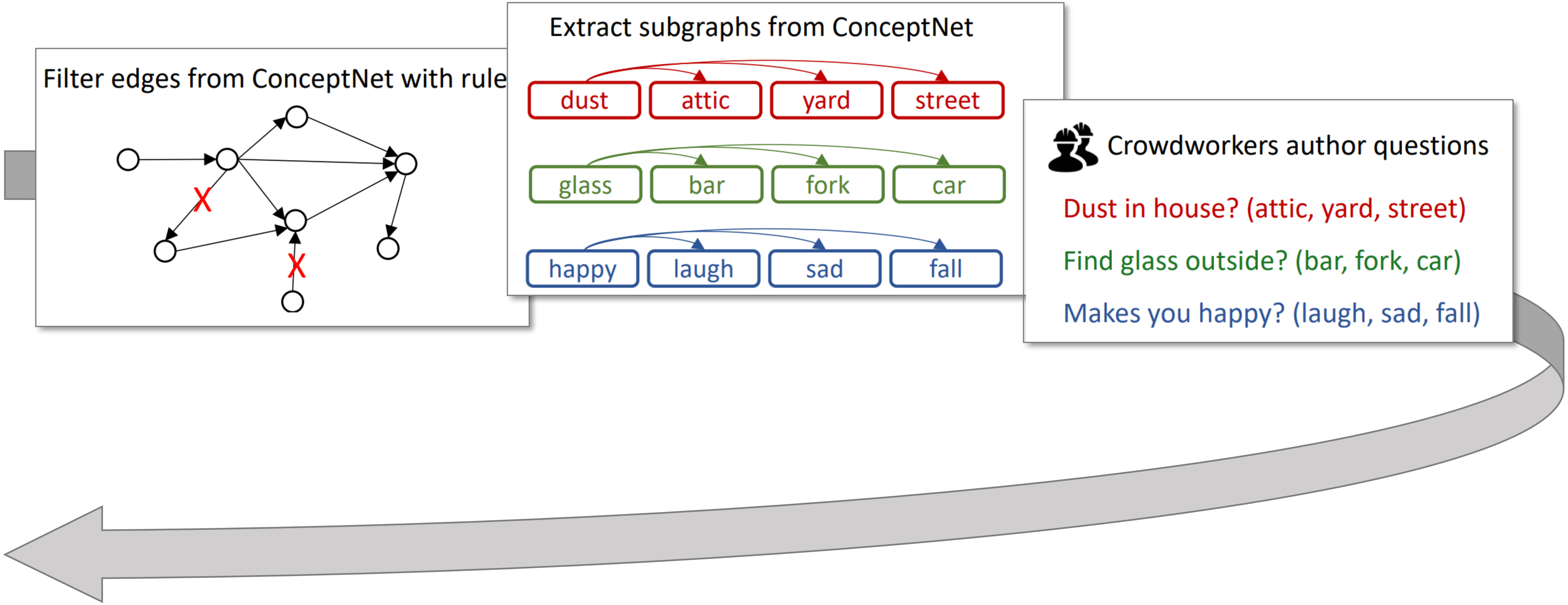
Talmor et al. 2019

COMMONSENSEQA: pivot on knowledge graphs



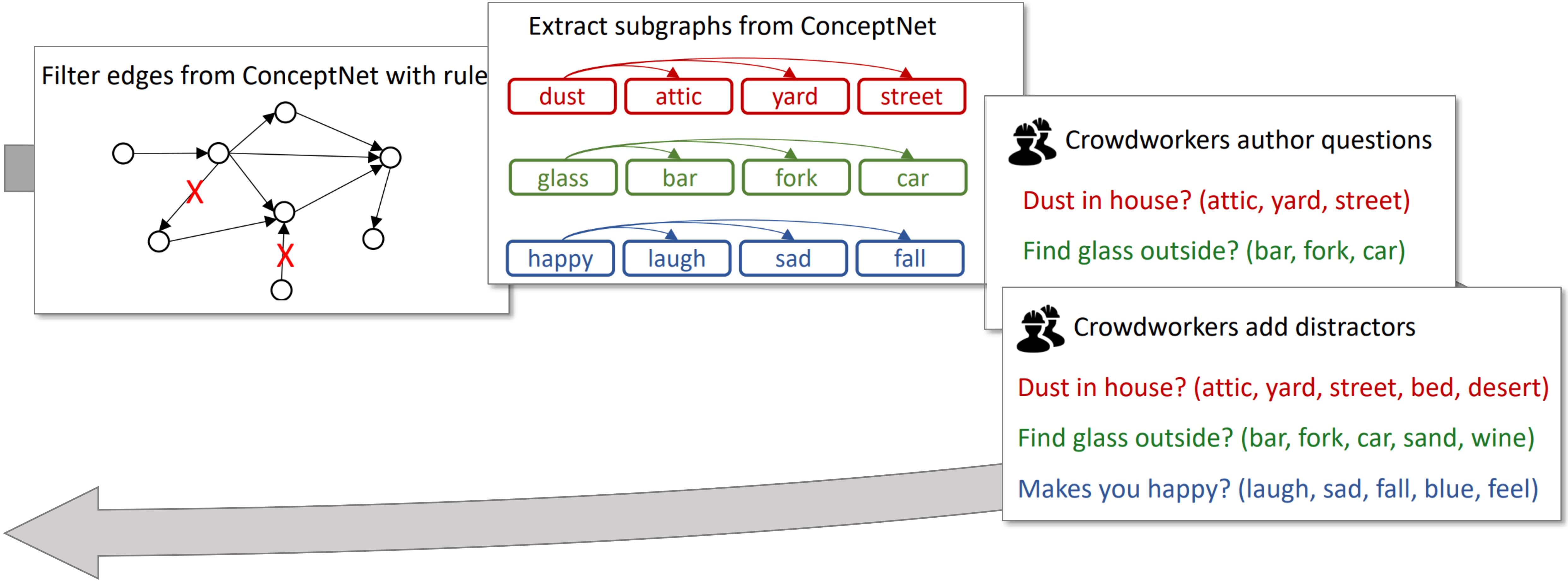
Talmor et al. 2019

CommonsenseQA: pivot on knowledge graphs



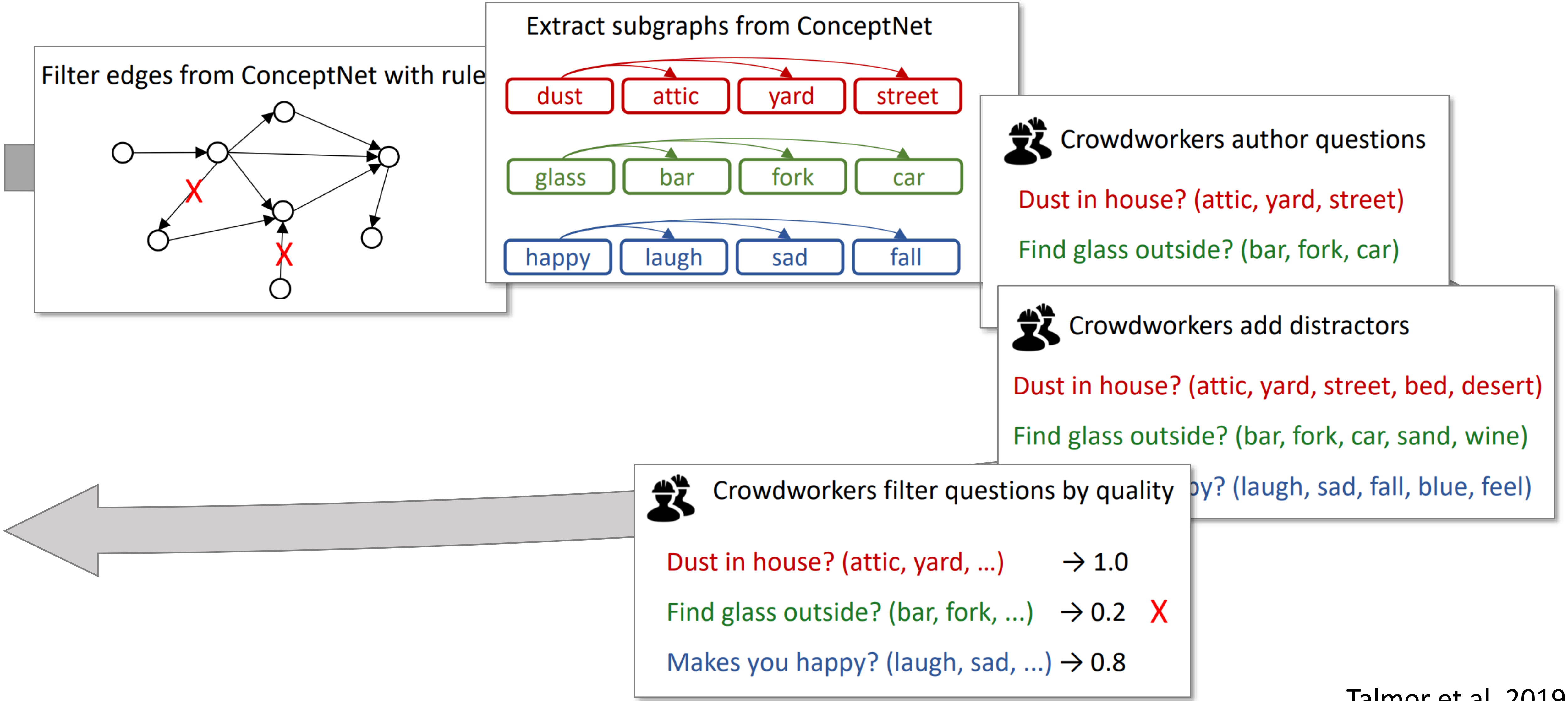
Talmor et al. 2019

COMMONSENSEQA: pivot on knowledge graphs



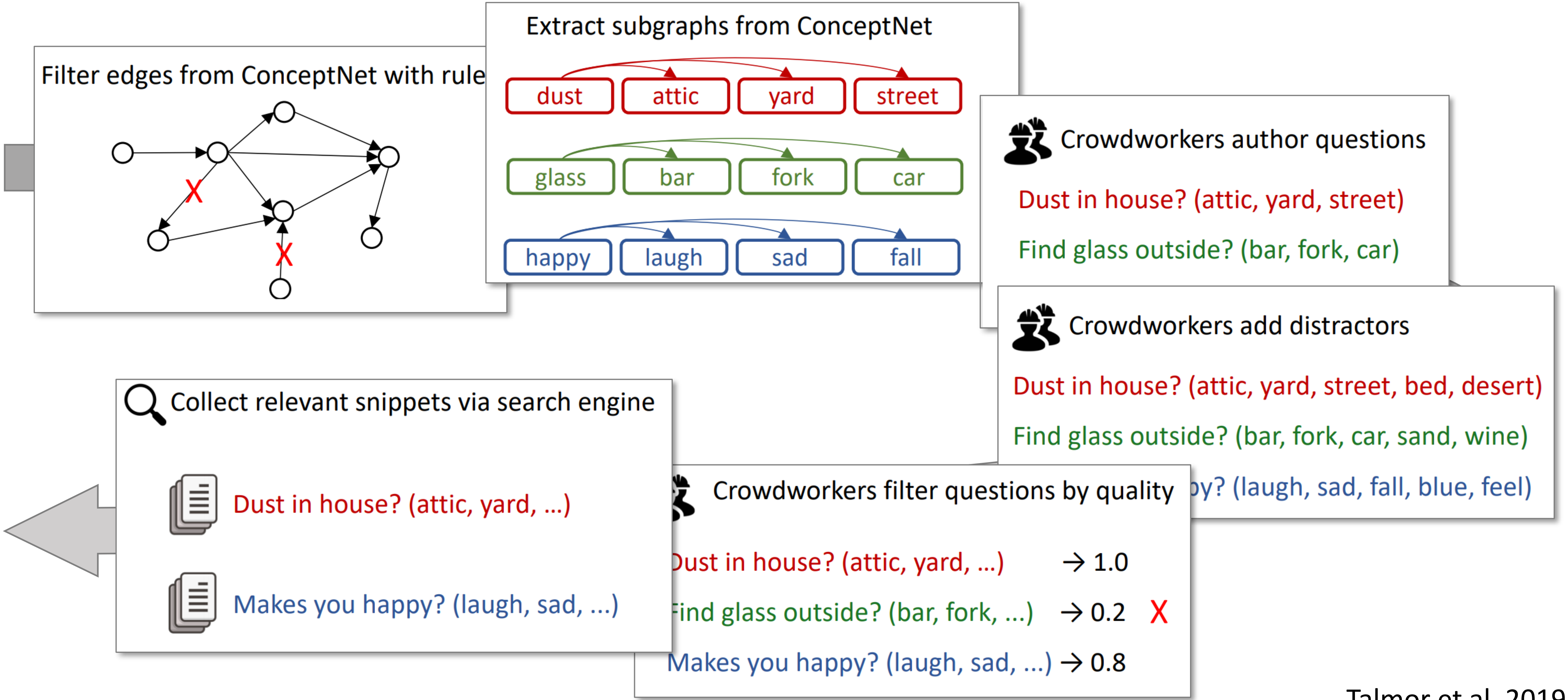
Talmor et al. 2019

COMMONSENSEQA: pivot on knowledge graphs



Talmor et al. 2019

COMMONSENSEQA: pivot on knowledge graphs



Talmor et al. 2019

Adversarial Filtering (lite)

Goal: remove examples with exploitable artifacts or spurious correlations

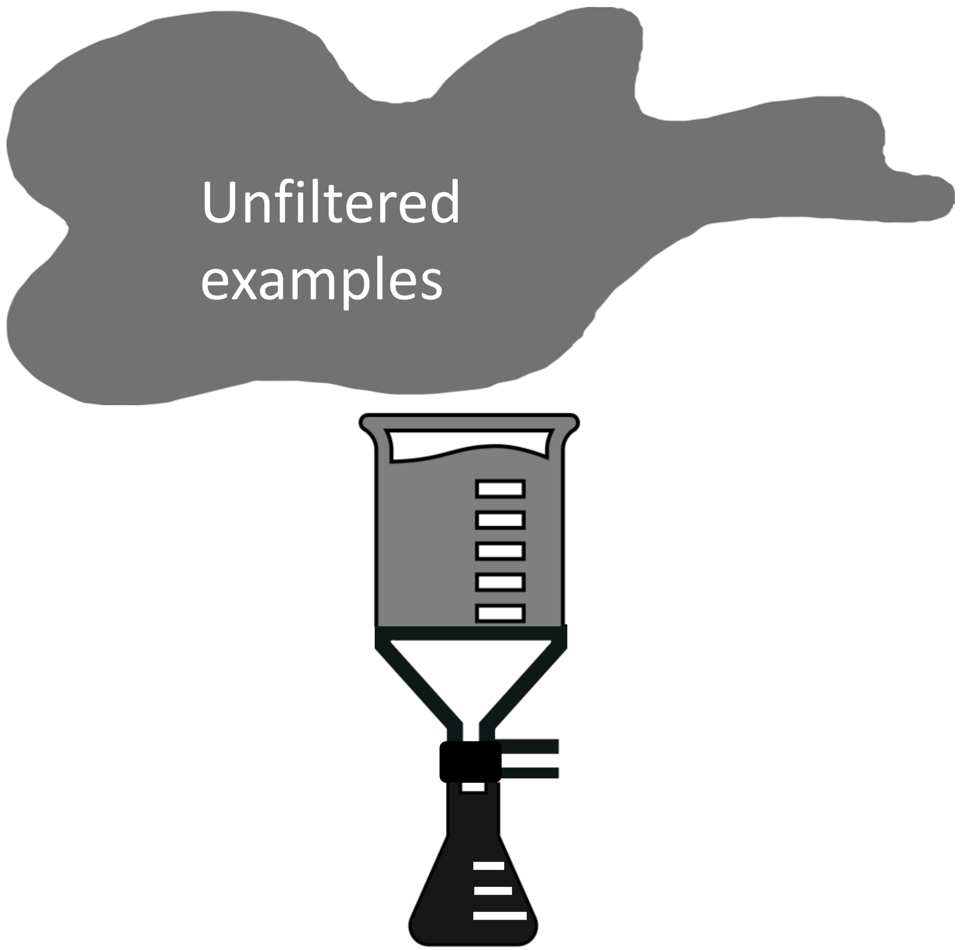
- Use pre-trained representations
- Iteratively remove data that's easiest to predict by a linear classifier (e.g., logistic)
- Robust examples remain

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

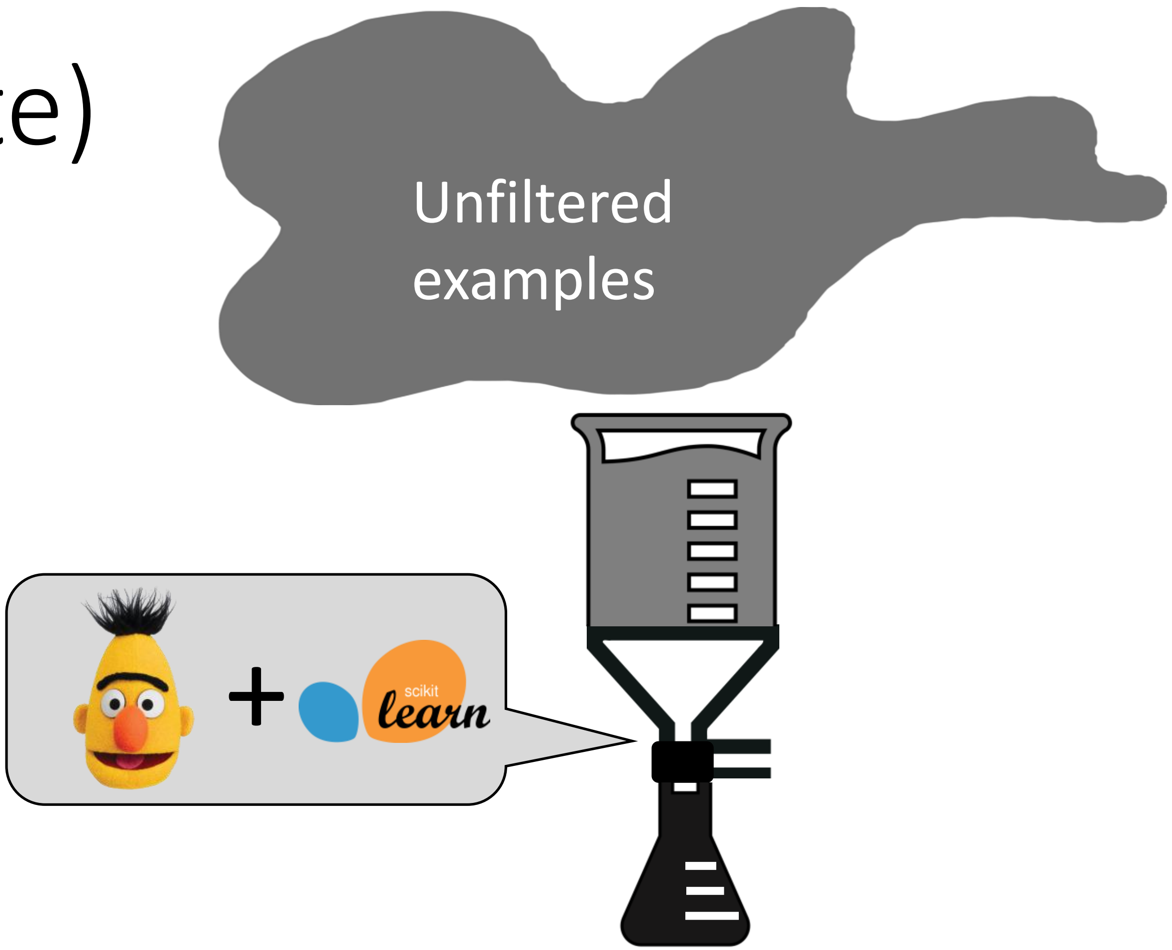


HellaSwag (Zellers et al., 2019)
AF-lite (Le Bras et al., 2019)

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

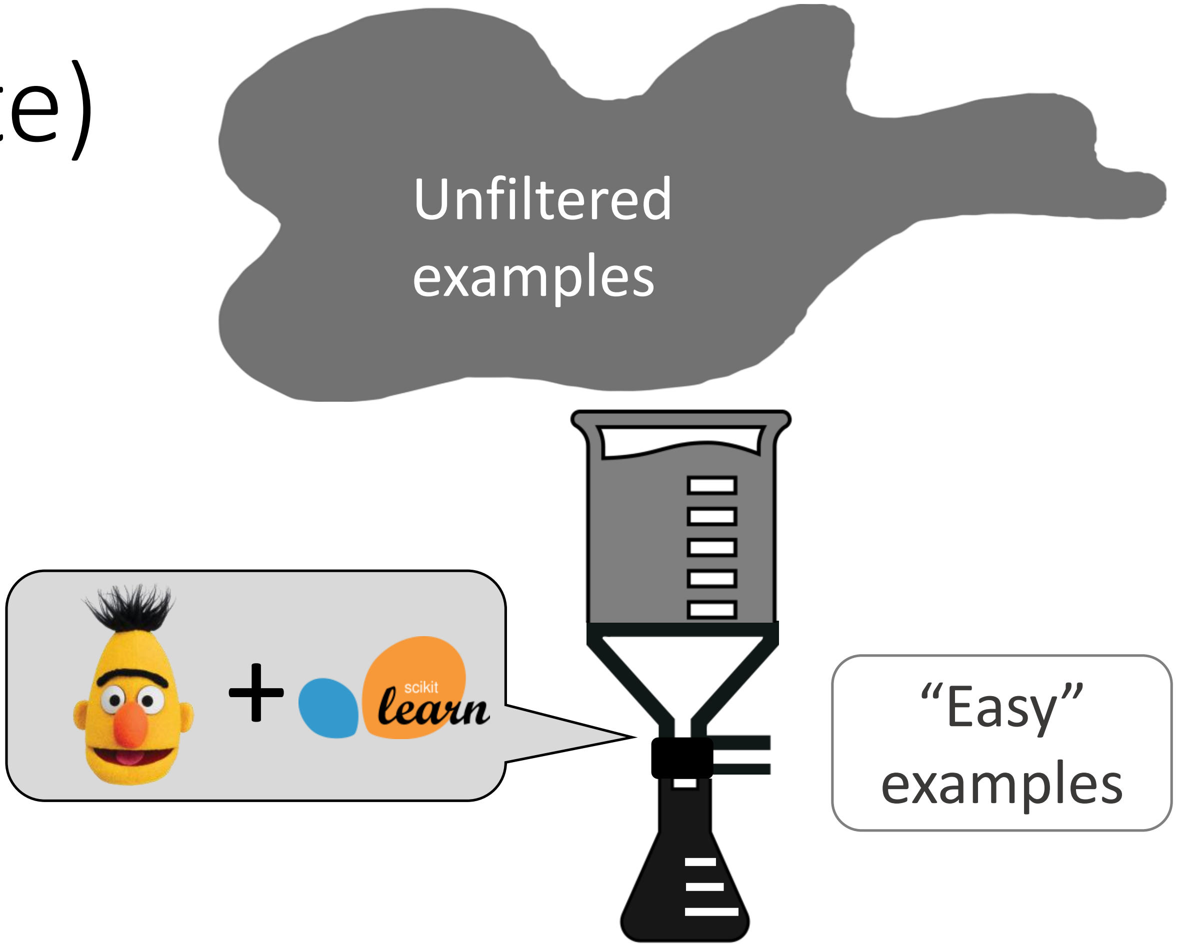


HellaSwag (Zellers et al., 2019)
AF-lite (Le Bras et al., 2019)

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

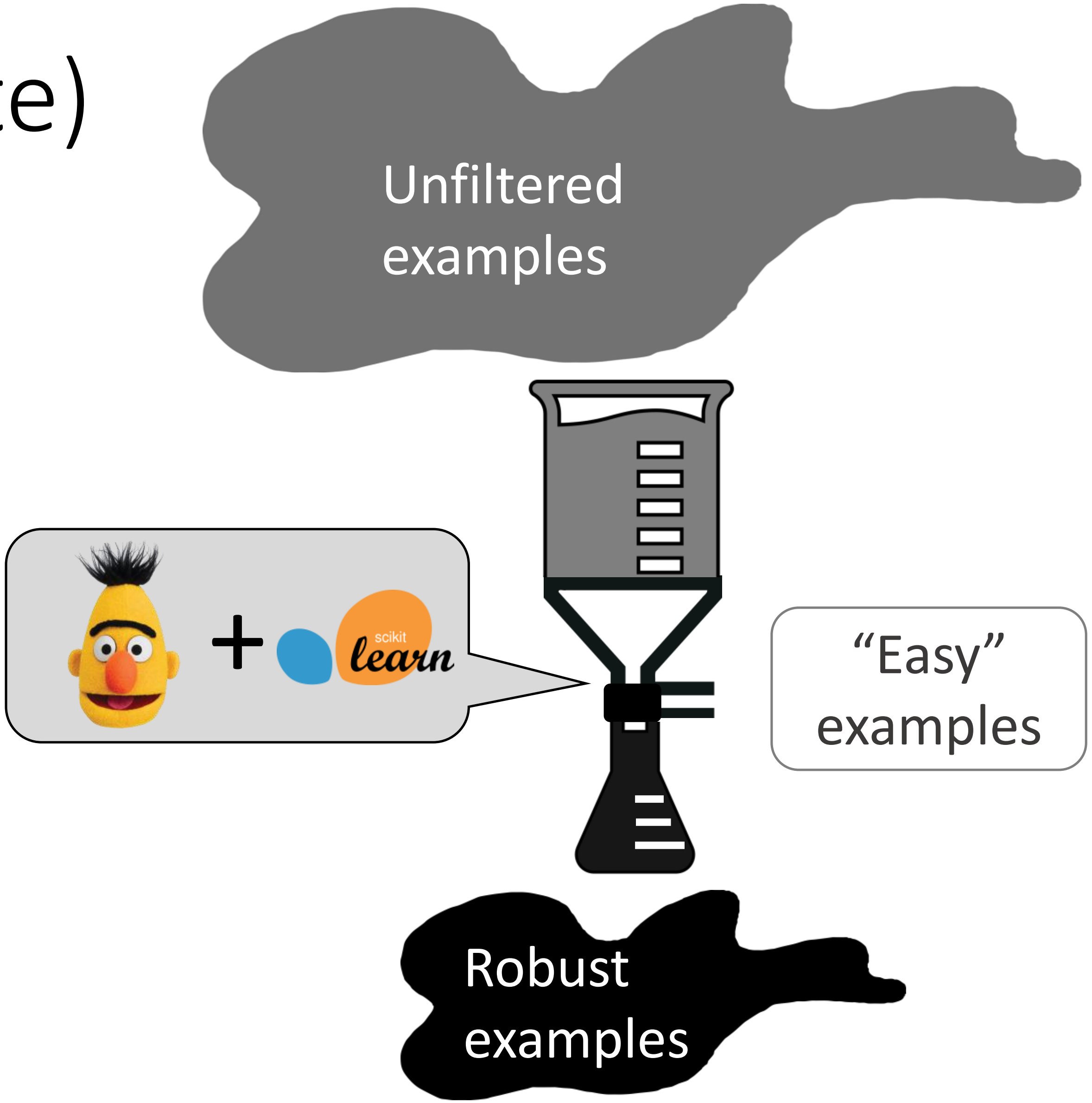


HellaSwag (Zellers et al., 2019)
AF-lite (Le Bras et al., 2019)

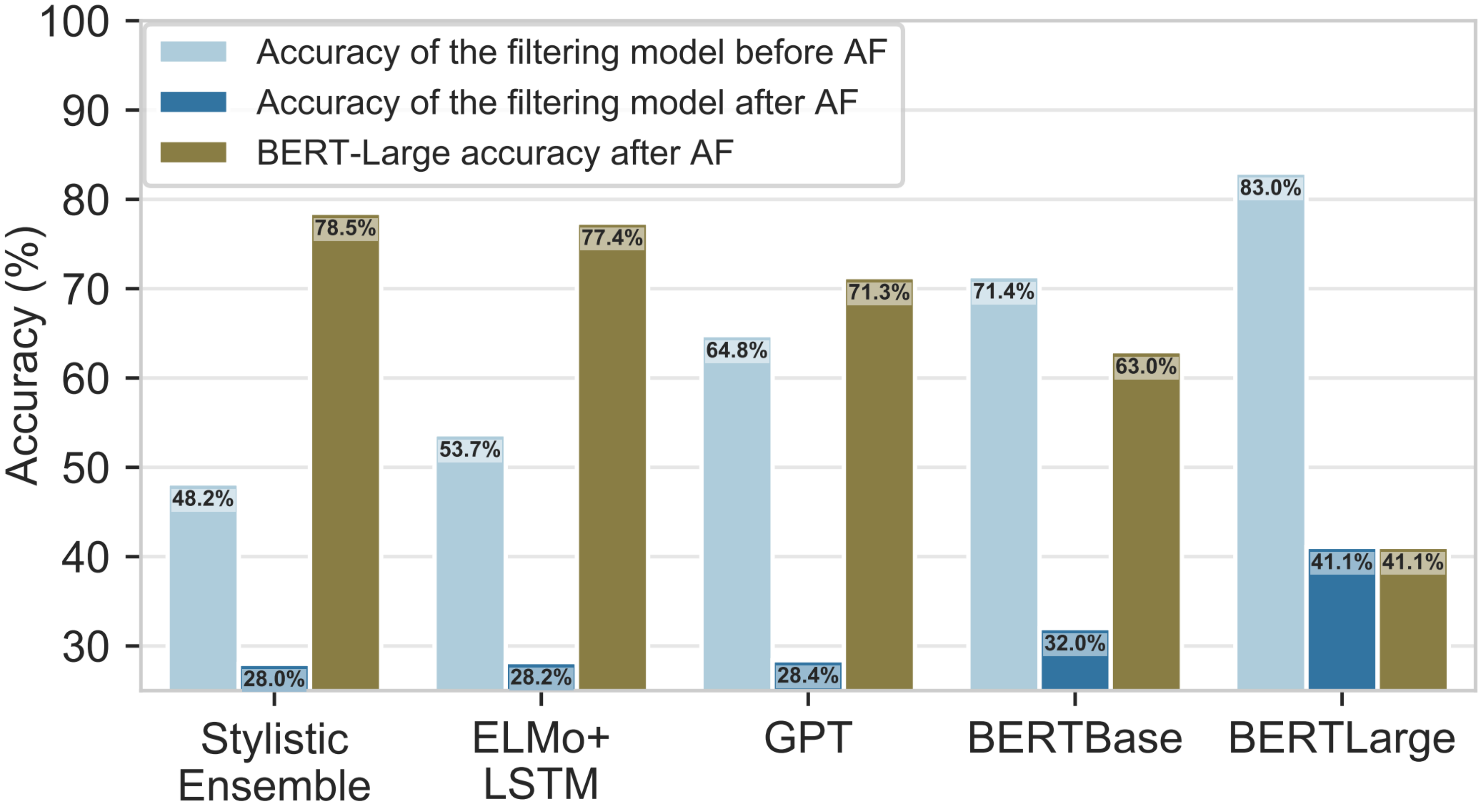
Adversarial Filtering (lite)

Goal: remove examples with exploitable artifacts or spurious correlations

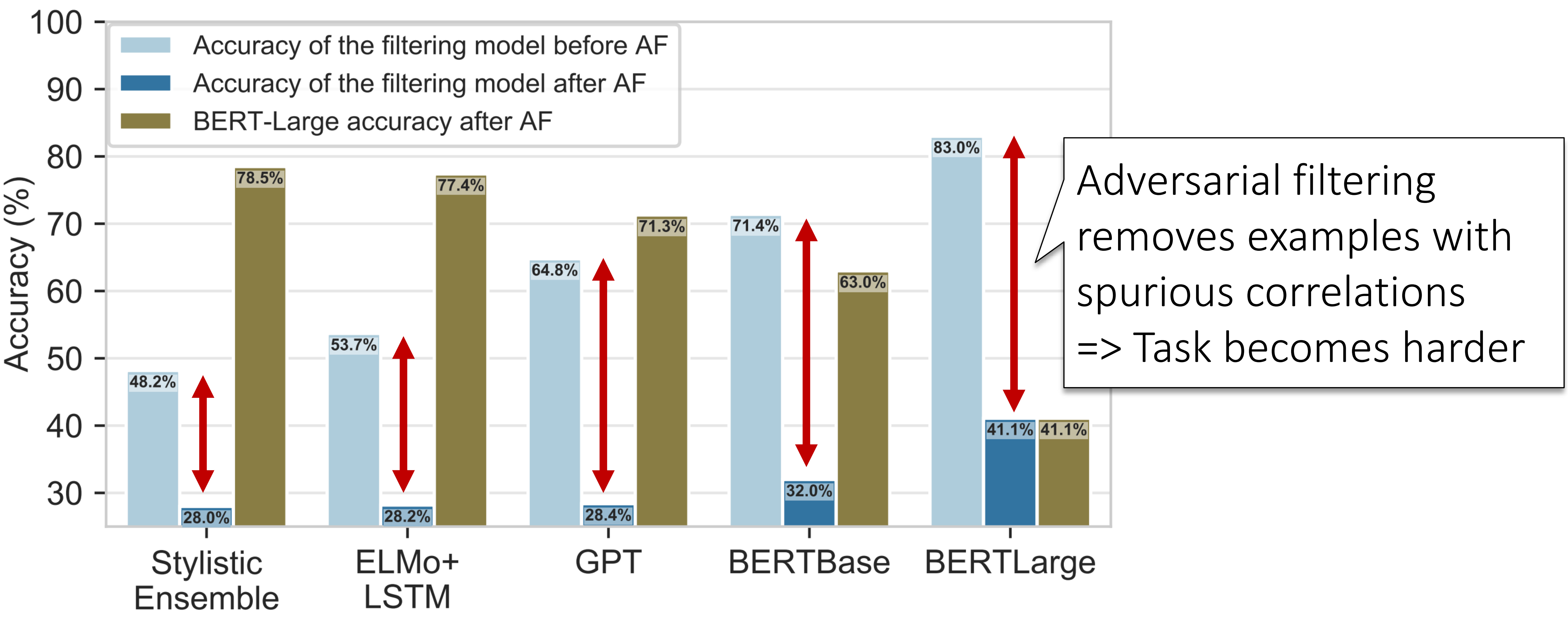
- Use pre-trained representations
- Iteratively remove data that's easiest to predict by a linear classifier (e.g., logistic)
- Robust examples remain



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

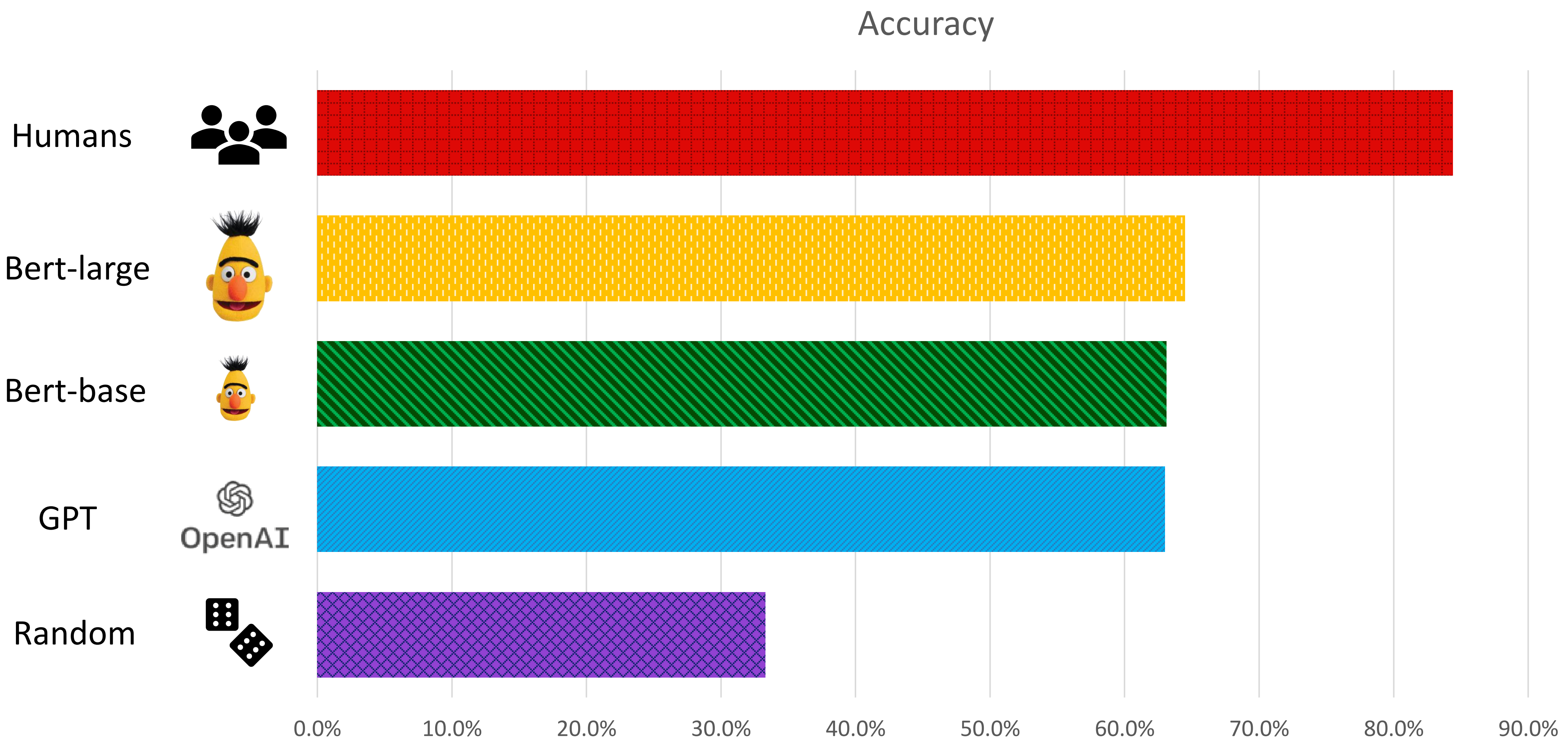


Performance of models on the WikiHow portion of HellaSwag (Zellers et al., 2019) with different AF settings and different training models

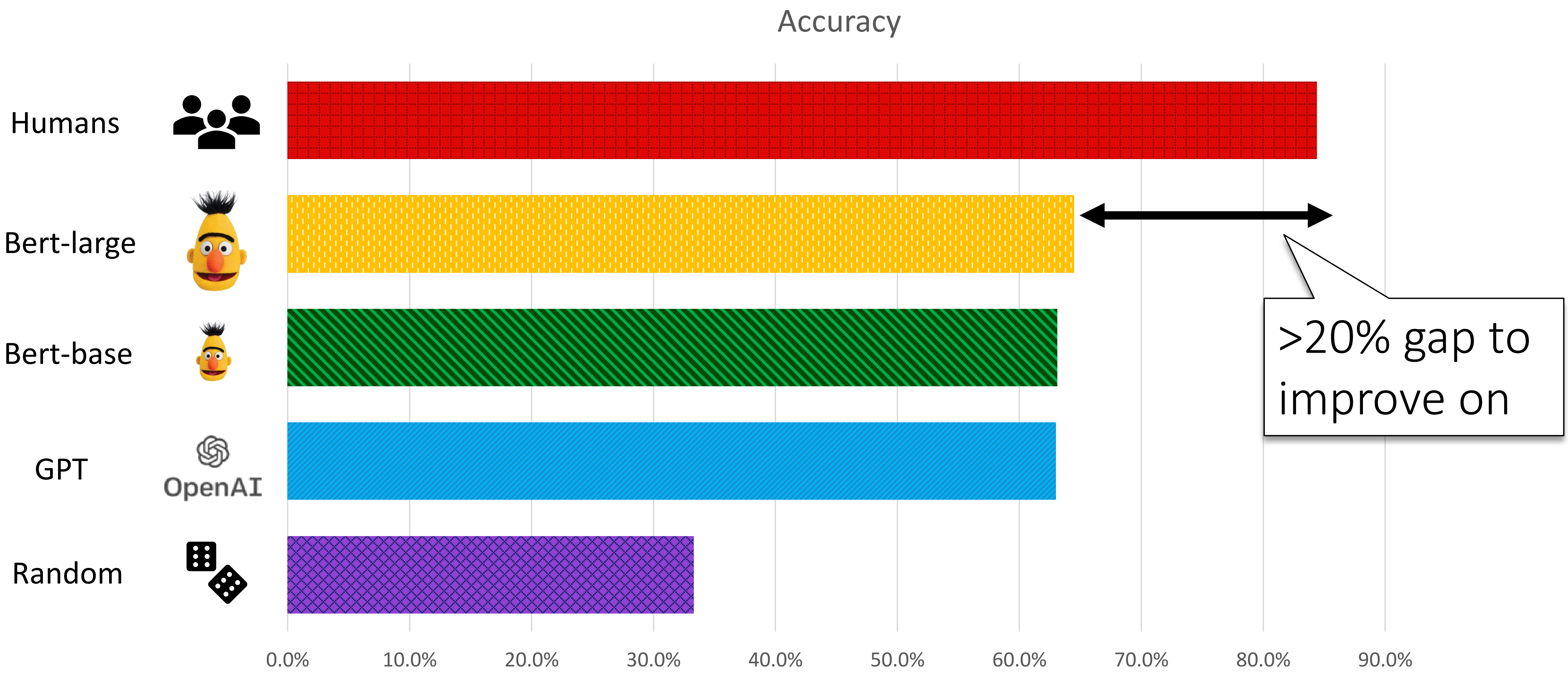


Performance of models on the WikiHow portion of HellaSwag (Zellers et al., 2019) with different AF settings and different training models

Model performance on SOCIAL IQA



Model performance on SOCIAL IQA



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 — how **Aubrey** would feel, not Alex

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 — how **Aubrey** would feel, not Alex
- ✓ boastful
- they need to practice more

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

What will Remy want to do next?

- lose her credit card
- arrive at a hotel — what Remy did **before**
- ✓ get the key from Skylar

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 — how **Aubrey** would feel, not Alex
- ✓ boastful
- they need to practice more

Need more robust, person-centric reasoning

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

What will Remy want to do next?

- lose her credit card
- arrive at a hotel — what Remy did **before**
- ✓ get the key from Skylar

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 — how **Aubrey** would feel, not Alex
- ✓ boastful
- they need to practice more

Need more robust, person-centric reasoning

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

What will Remy want to do next?

- lose her credit card
- arrive at a hotel — what Remy did **before**
- ✓ get the key from Skylar

Need better notion of causes vs. effects

Commonsense benchmarks

Naïve Psychology

ROC story

Social IQa

Physical IQa

HellaSwag

SWAG

Abductive NLI

WSC

COPA

VCR

WinoGrande

CommonsenseQA

JHU Ordinal Commonsense

MCTaco

ReCORD

CosmosQA

MultiRC

Commonsense benchmarks

Social commonsense

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

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

- MCTaco

- ReCORD
- CosmosQA
- MultiRC

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

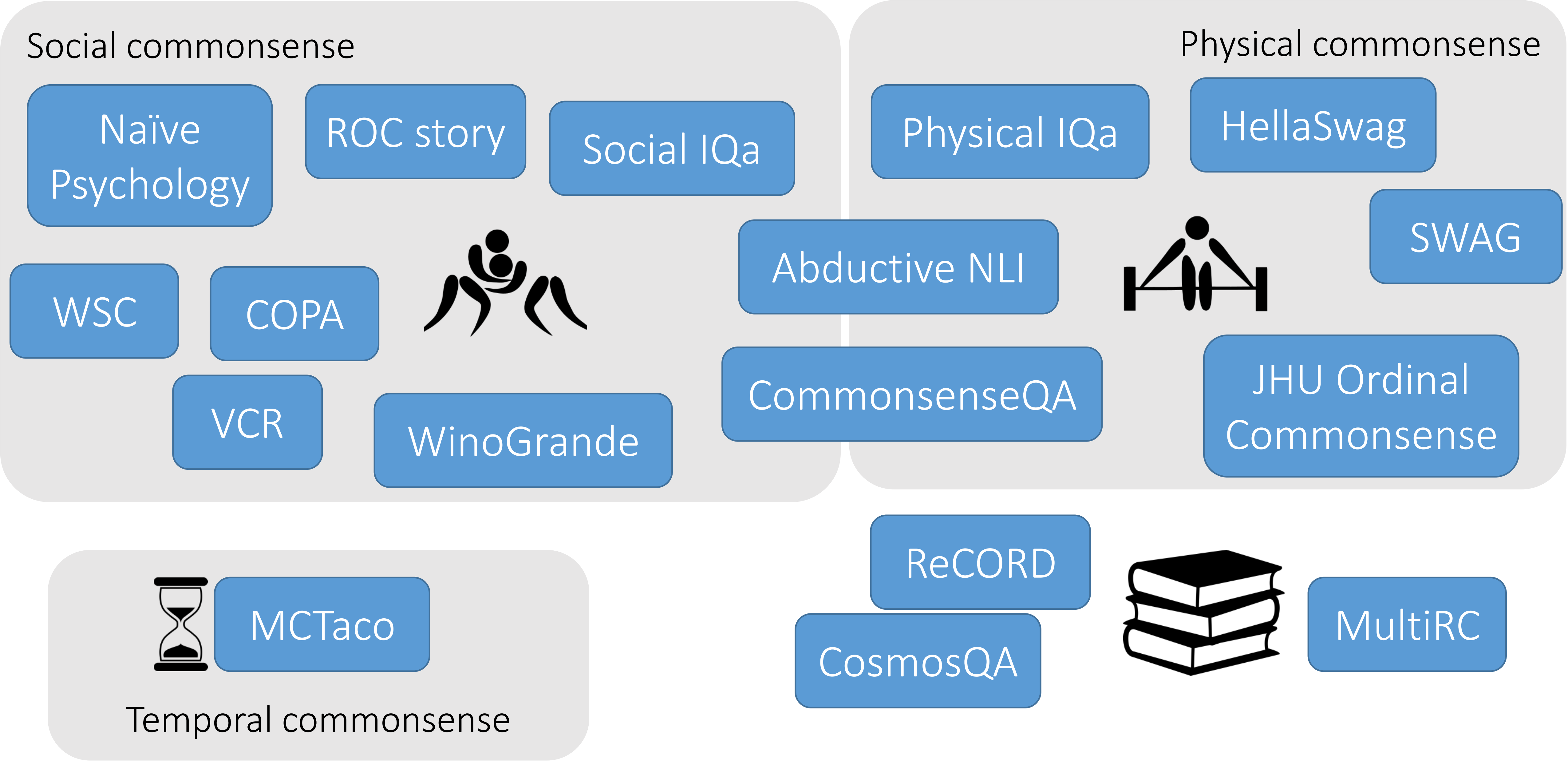
MCTaco

ReCORD

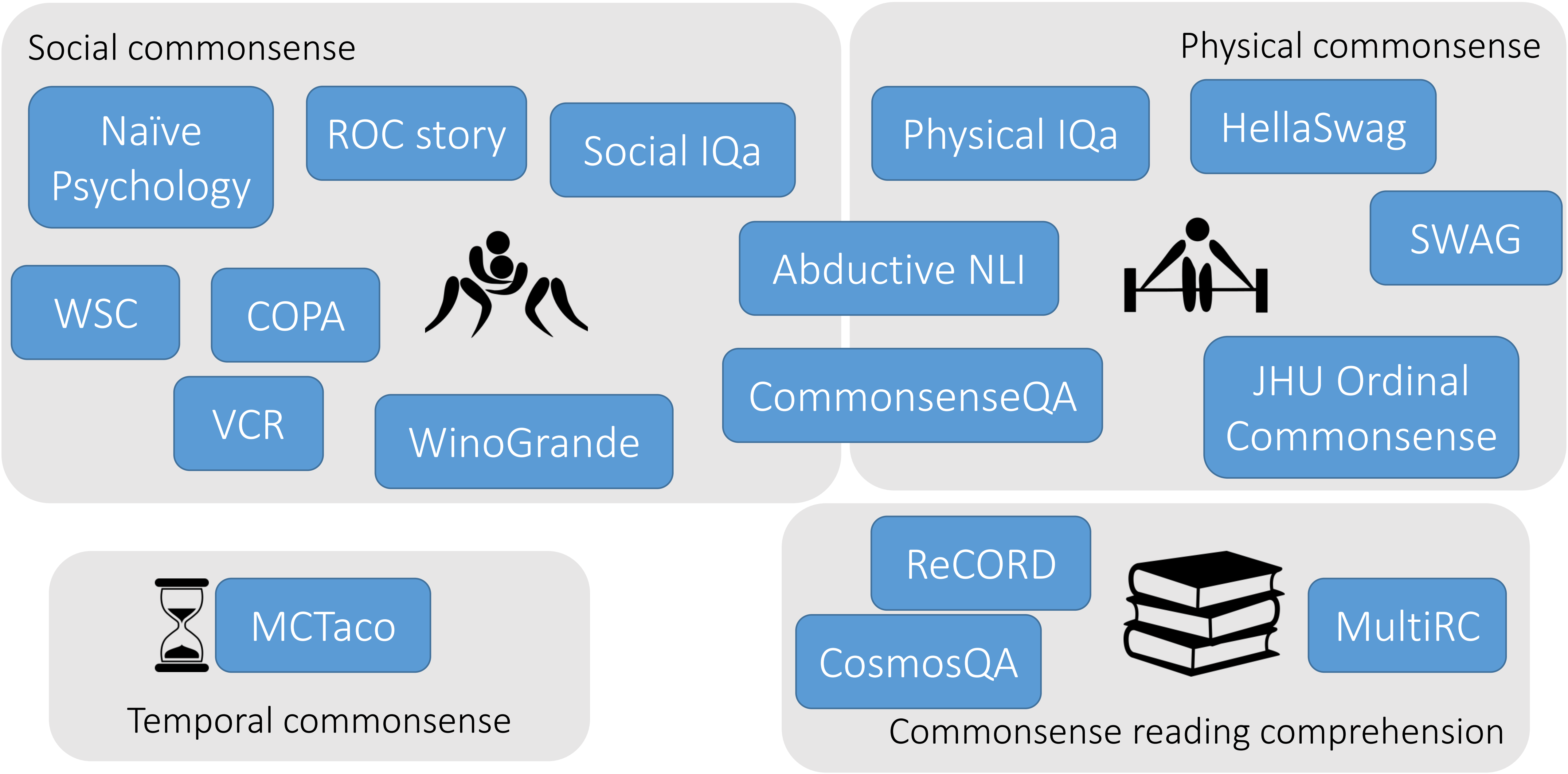
CosmosQA

MultiRC

Commonsense benchmarks



Commonsense benchmarks



Commonsense benchmarks

Social commonsense

Naïve Psychology

ROC story

Social IQa

Physical commonsense

Physical IQa

HellaSwag

SWAG

Abductive NLI

WSC

Thanks! Questions?

VCR

WinoGrande

CommonsenseQA

JHU Ordinal Commonsense

MCTaco

Temporal commonsense

ReCORD

CosmosQA

MultiRC

Commonsense reading comprehension