Using Commonsense Knowledge to Answer Why-Questions

Yash Kumar Lal

Niket Tandon

Tanvi Aggarwal

Horace Liu

Nate Chambers

Raymond Mooney

Niranjan Balasubramanian

Progress on many types of QA problems

Reading Comprehension

Knowledge Graphs

Tabular Data

Conversational QA

…

Rajpurkar & Jia et al. ‘18; Yu et al. ‘18; Zhu et al. ‘21; Reddy et al. ‘19

But What About Commonsense Based QA in Narratives?

Matt and Sarah were pregnant. They wanted to announce it in a fun way. They wrote it on a cake. They invited their friends over. When their friends saw the cake, they were excited.

Mostafazadeh et al. ‘16

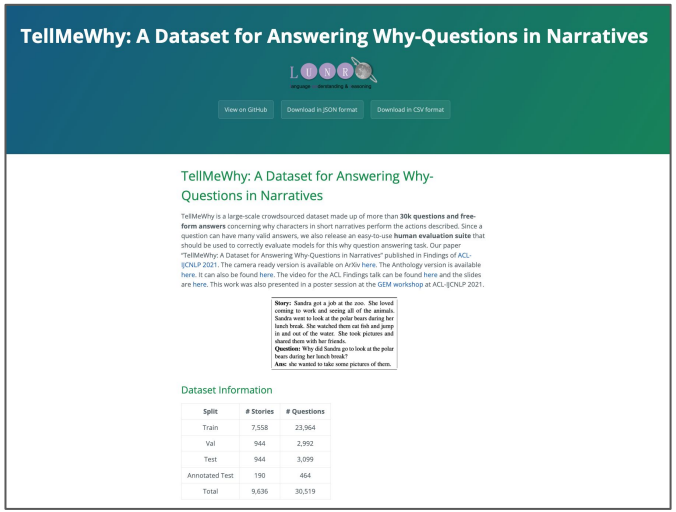
But What About Commonsense Based QA in Narratives?

Matt and Sarah were pregnant. They wanted to announce it in a fun way. They wrote it on a cake. They invited their friends over. When their friends saw the cake, they were excited.

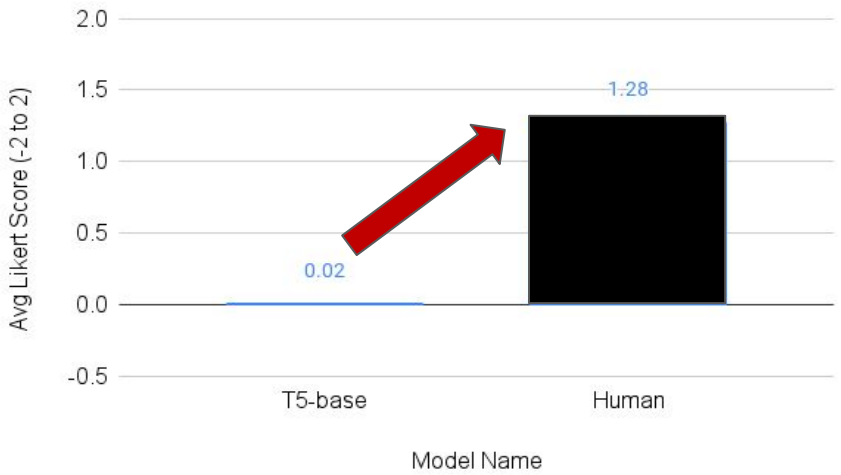
Q: Why were Matt and Sarah pregnant?

They wanted to have a baby.

Why Question Answering with TellMeWhy



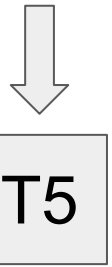
[Lal et al., ACL-Findings 2021]



How Can We Improve Why QA Performance?

Matt and Sarah were pregnant. They wanted to announce it in a fun way. They wrote it on a cake. They invited their friends over. When their friends saw the cake, they were excited.

Q: Why were Matt and Sarah pregnant?



How Can We Improve Why QA Performance?

Matt and Sarah were pregnant. They wanted to announce it in a fun way. They wrote it on a cake. They invited their friends over. When their friends saw the cake, they were excited.

Q: Why were Matt and Sarah pregnant?

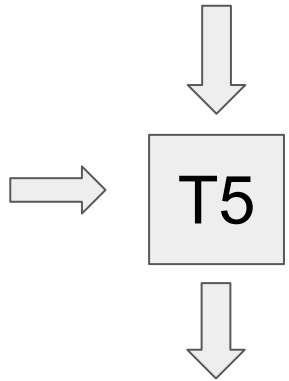
Matt and Sarah were pregnant

How Can We Improve Why QA Performance?

Matt and Sarah were pregnant. They wanted to announce it in a fun way. They wrote it on a cake. They invited their friends over. When their friends saw the cake, they were excited.

Q: Why were Matt and Sarah pregnant?

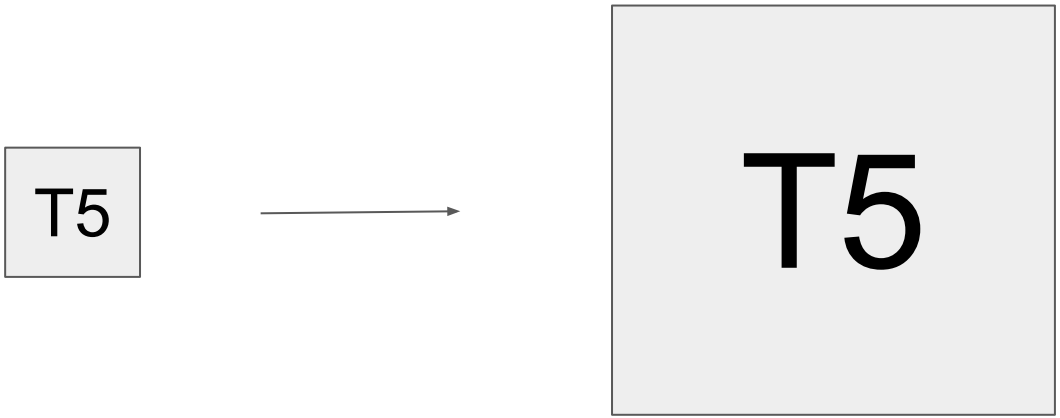
Commonsense Knowledge:
- ❑ become pregnant to have babies
- ❑ can become pregnant from sexual intercourse



They wanted to have a baby

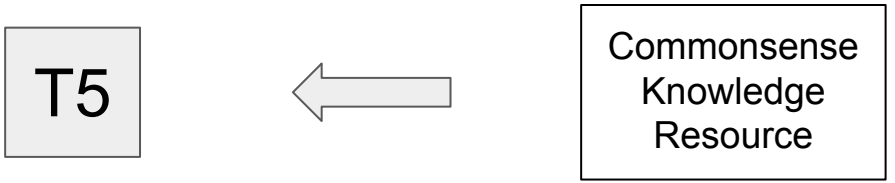
How Can We Improve Access to Commonsense Knowledge

Larger Models?



How Can We Improve Access to Commonsense Knowledge

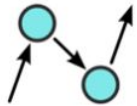
External Knowledge?



How Can We Improve Access to Commonsense Knowledge

External Knowledge?

T5



ConceptNet
An open, multilingual knowledge graph

GLUCOSE Dataset

COMET

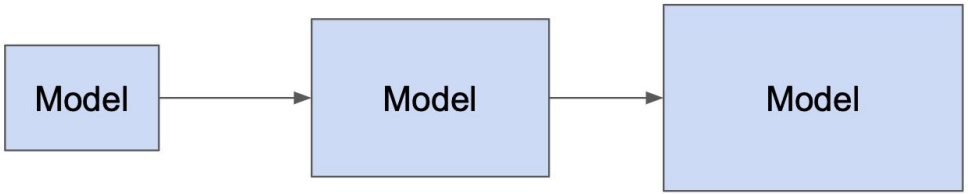
Speer et al. ‘17; Mostafazadeh et al. ‘20; Hwang et al. ‘21

How Can We Improve Access to Commonsense Knowledge

External Knowledge?

Our Contributions

Larger Models

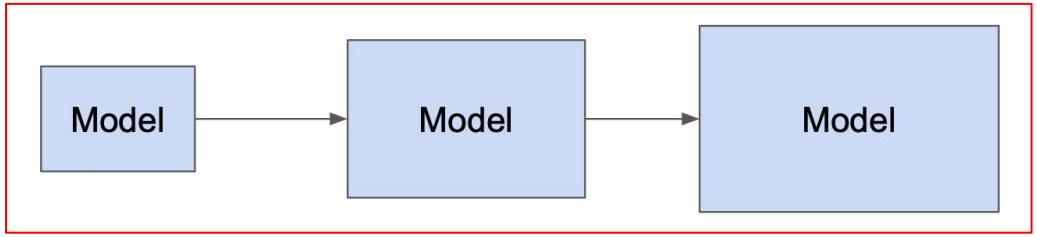


External Knowledge

COMET

Our Contributions

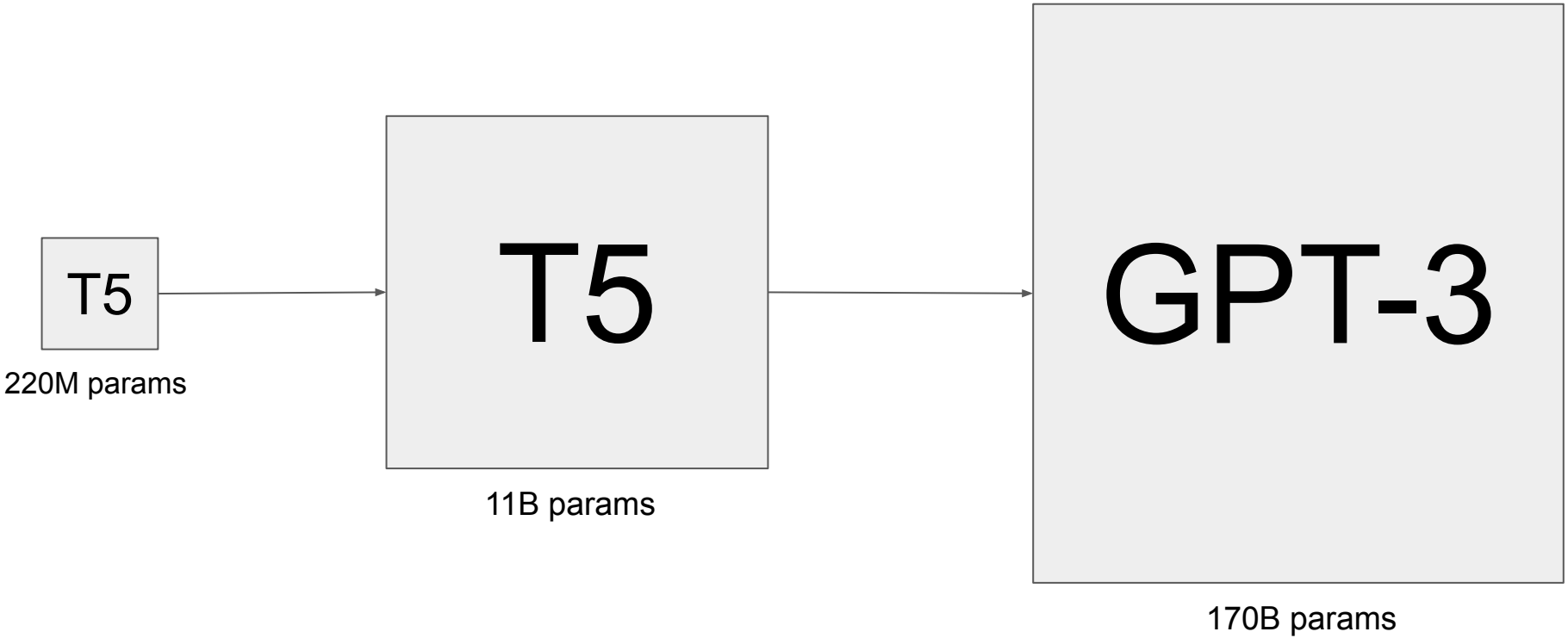
Larger Models



External Knowledge

COMET

How Far Can Larger Models Take Us?



Evaluation Setup

Q: Why were Matt and Sarah pregnant?

Model Answer: Matt and Sarah were pregnant.

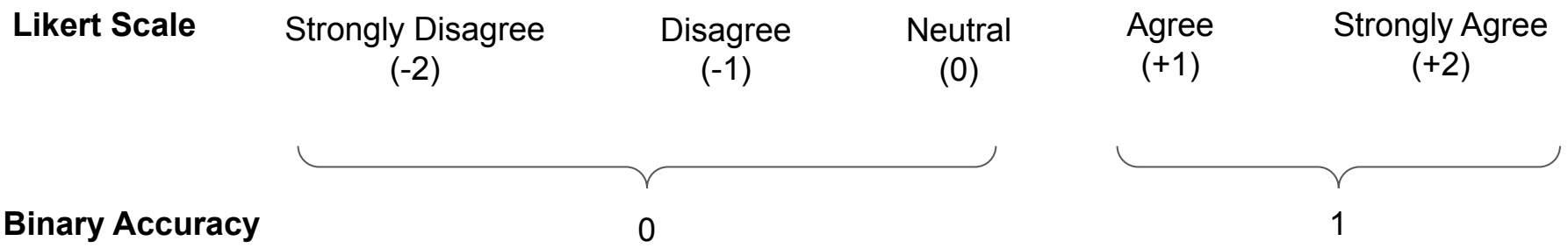
The answer is valid and makes sense given the story.

Evaluation Setup

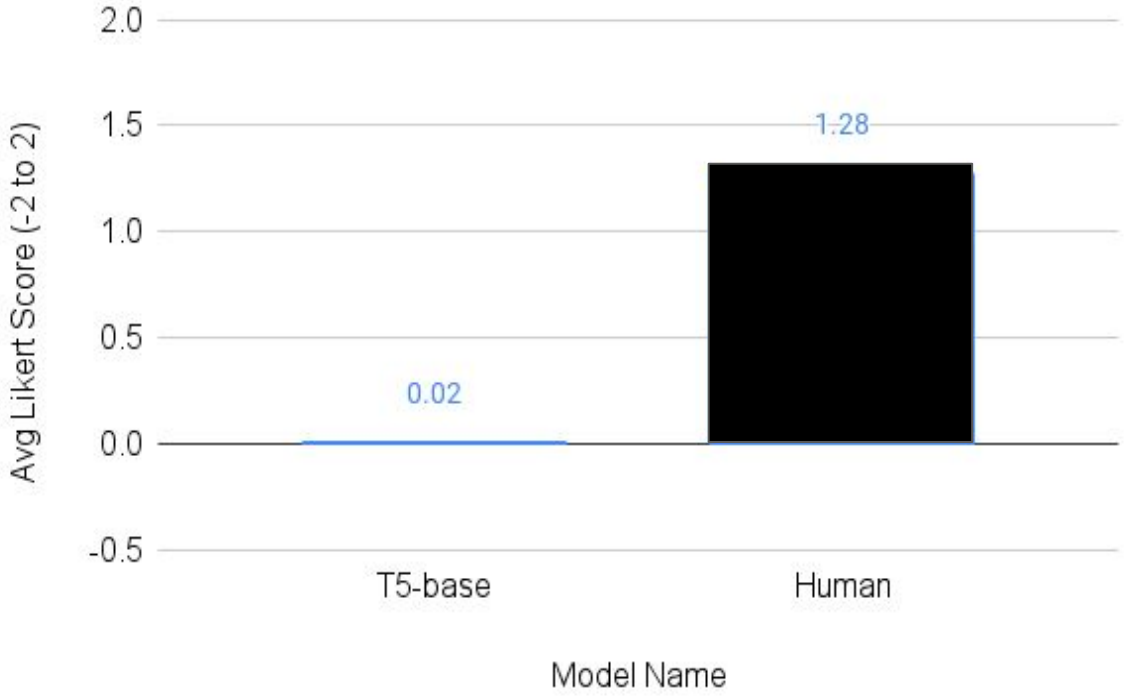
Likert Scale	Strongly Disagree (-2)	Disagree (-1)	Neutral (0)	Agree (+1)	Strongly Agree (+2)

Average Likert Score

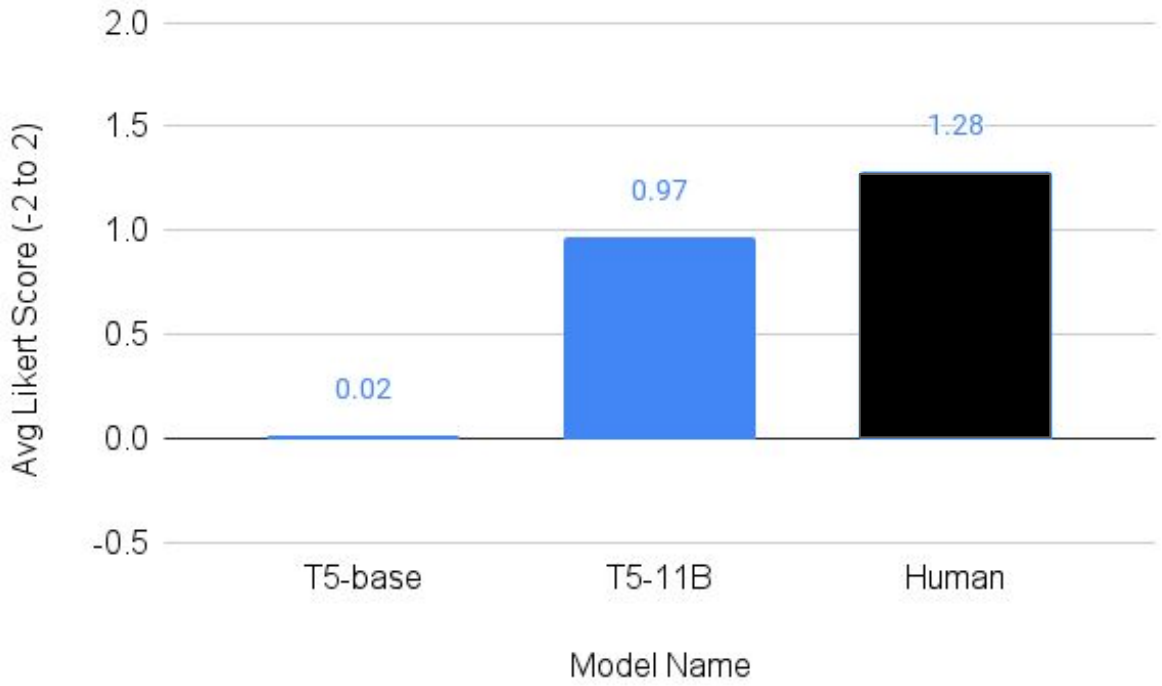
Evaluation Setup



Using Larger Models

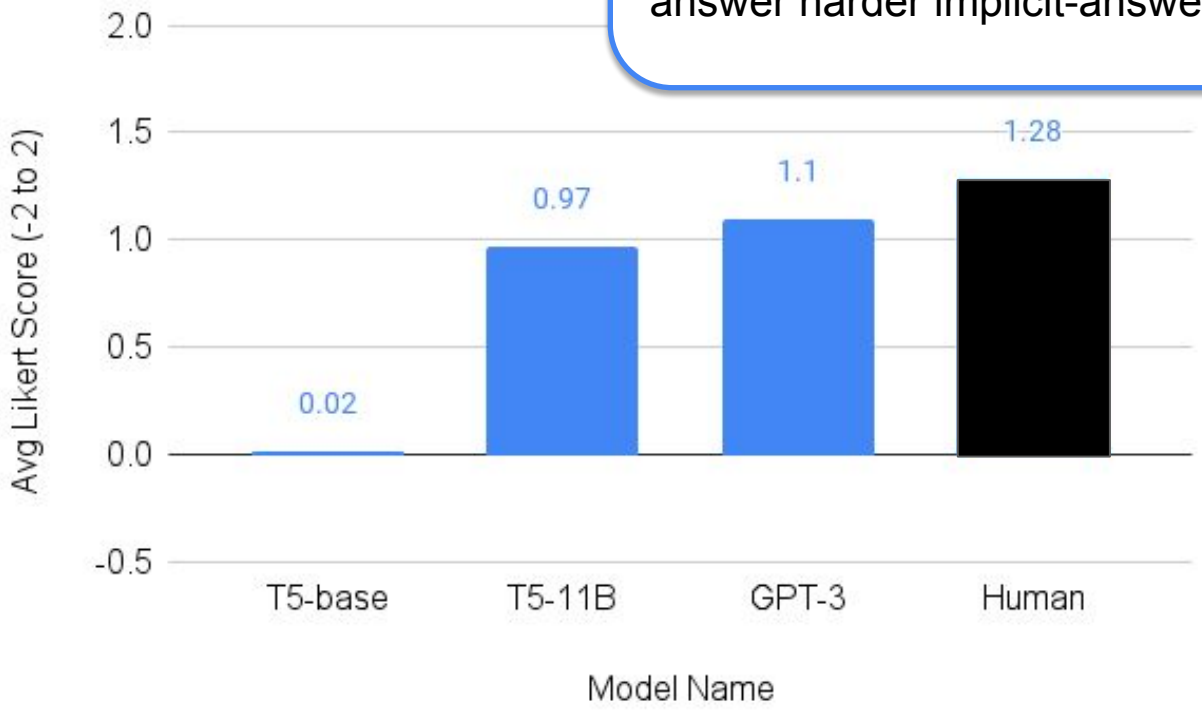


Using Larger Models



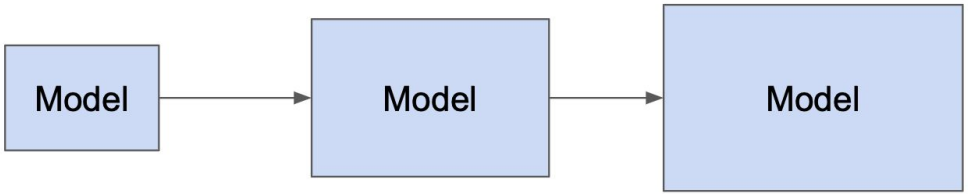
Using Larger Models

Larger models do seem to exhibit aspects of commonsense knowledge that allow them to answer harder implicit-answer questions.



Our Contributions

Larger Models



External Knowledge

COMET

COMET

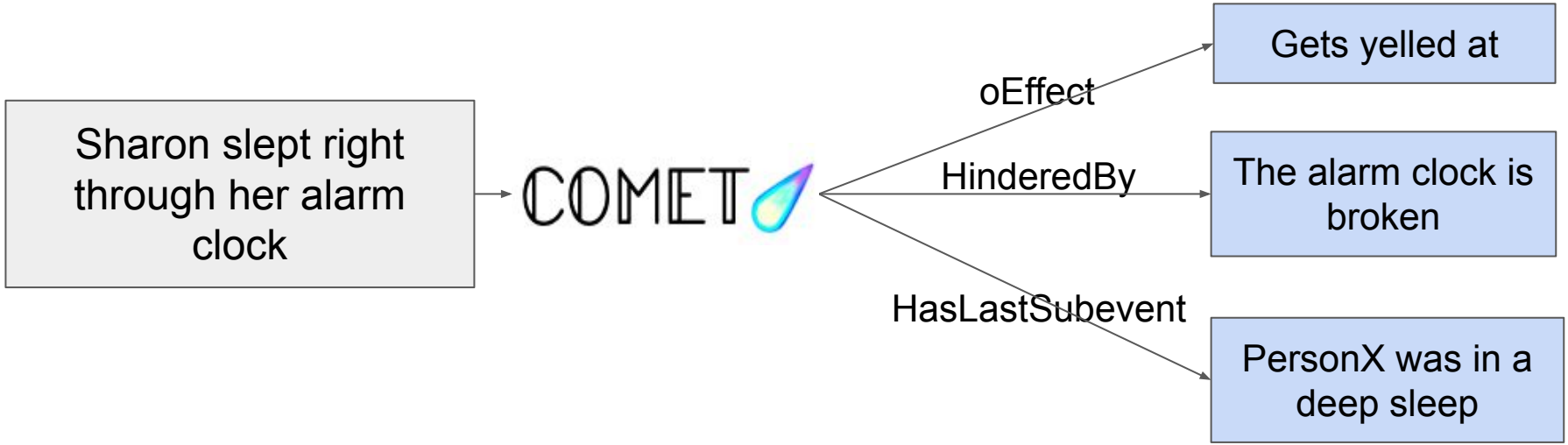
Sharon slept right through her alarm clock! She was dreaming of the ocean and how the waves sounded. She didn’t even hear her family leave for church. When she woke up and everyone was gone, she was afraid. She couldn’t believe they had gone without her.

COMET

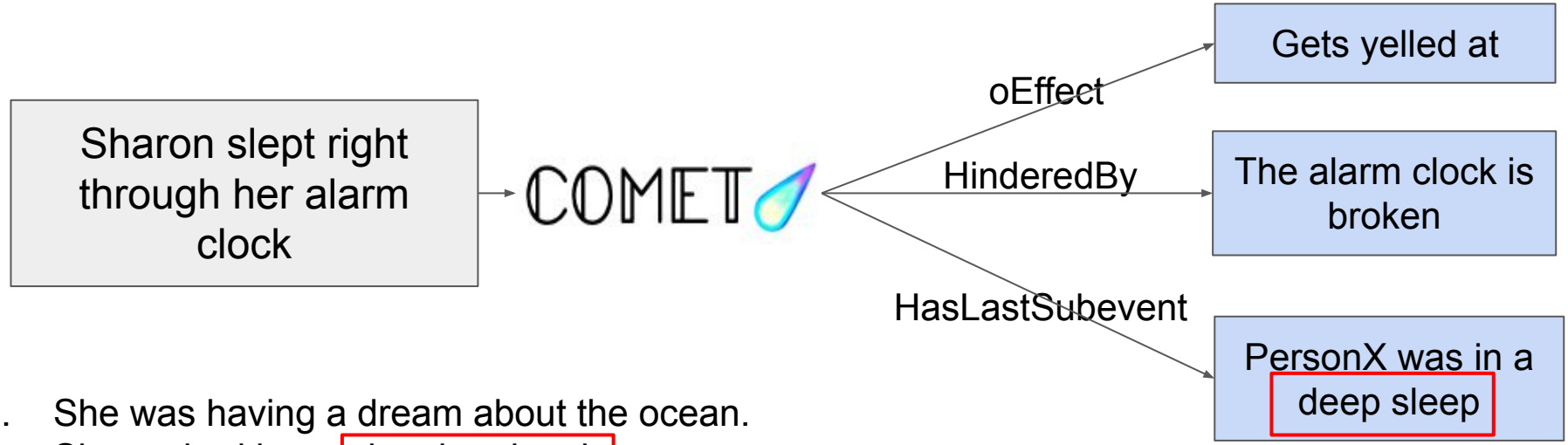
Sharon slept right through her alarm clock! She was dreaming of the ocean and how the waves sounded. She didn't even hear her family leave for church. When she woke up and everyone was gone, she was afraid. She couldn't believe they had gone without her.

Q: Why did Sharon sleep right through her alarm clock?

COMET


Sharon slept right through her alarm clock
COMET
oEffect
HinderedBy
HasLastSubevent
Gets yelled at
The alarm clock is broken
PersonX was in a deep sleep

COMET



1. She was having a dream about the ocean.
2. Sharon had been sleeping deeply.
3. She was in a deep sleep and dreaming of wave sounds.

Using COMET

Knowledge Selection

Not all knowledge from COMET is relevant

Amount of Knowledge

Might depend on model size

Knowledge Format

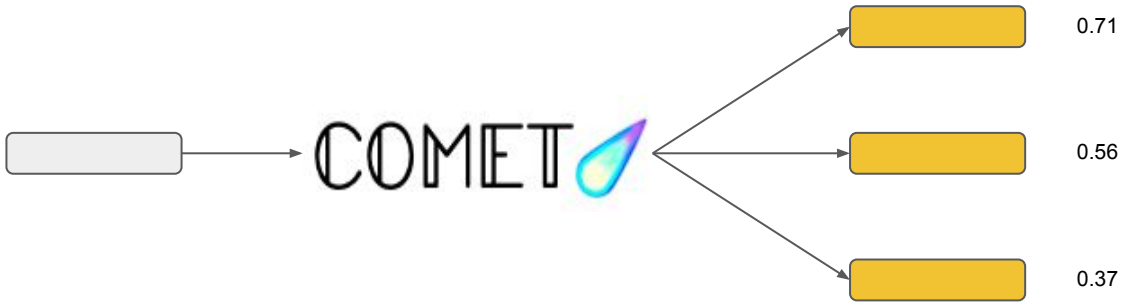
Models are known to be sensitive to input structure

Knowledge Selection

- COMET scores
- Trained re-ranking model
- Re-ranking + diversity metric

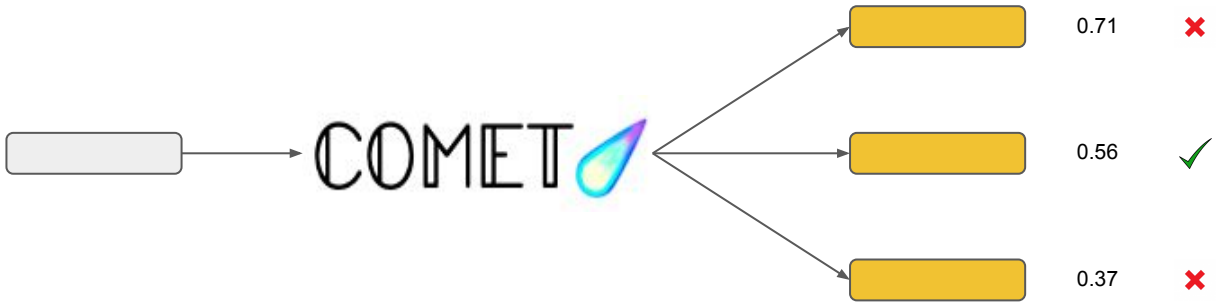
Knowledge Selection

- COMET scores



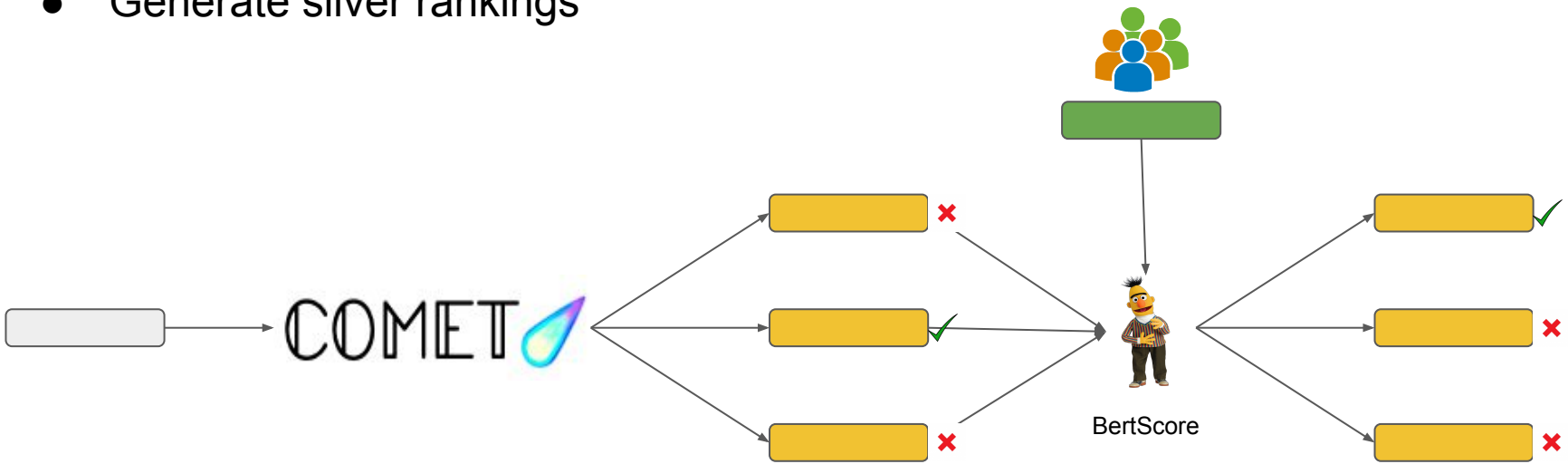
Knowledge Selection

- Trained re-ranking model



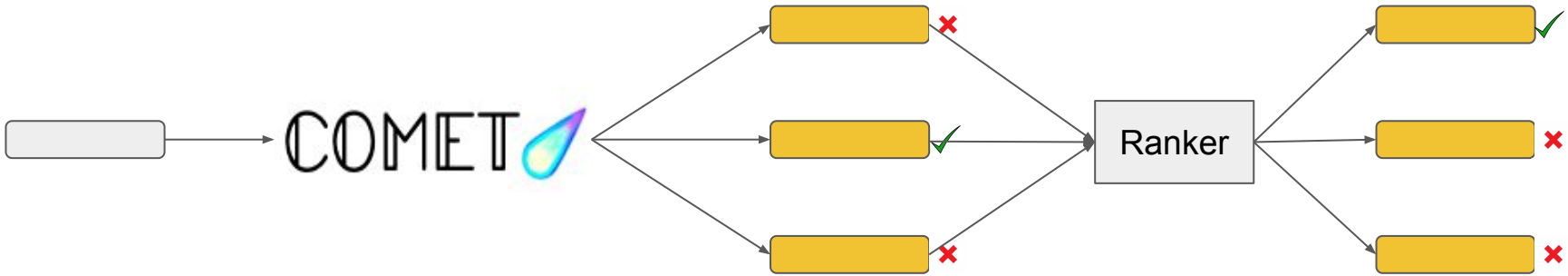
Knowledge Selection

- Generate silver rankings



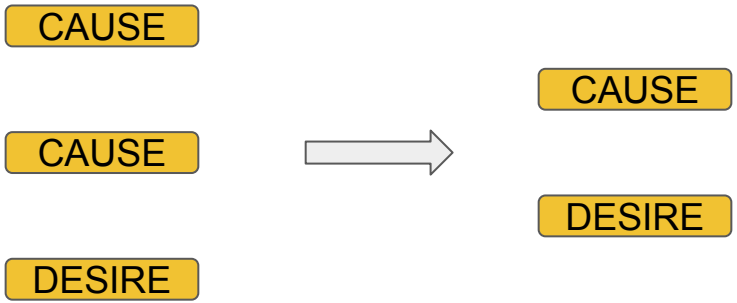
Knowledge Selection

- Generate silver rankings
- Train a re-ranker

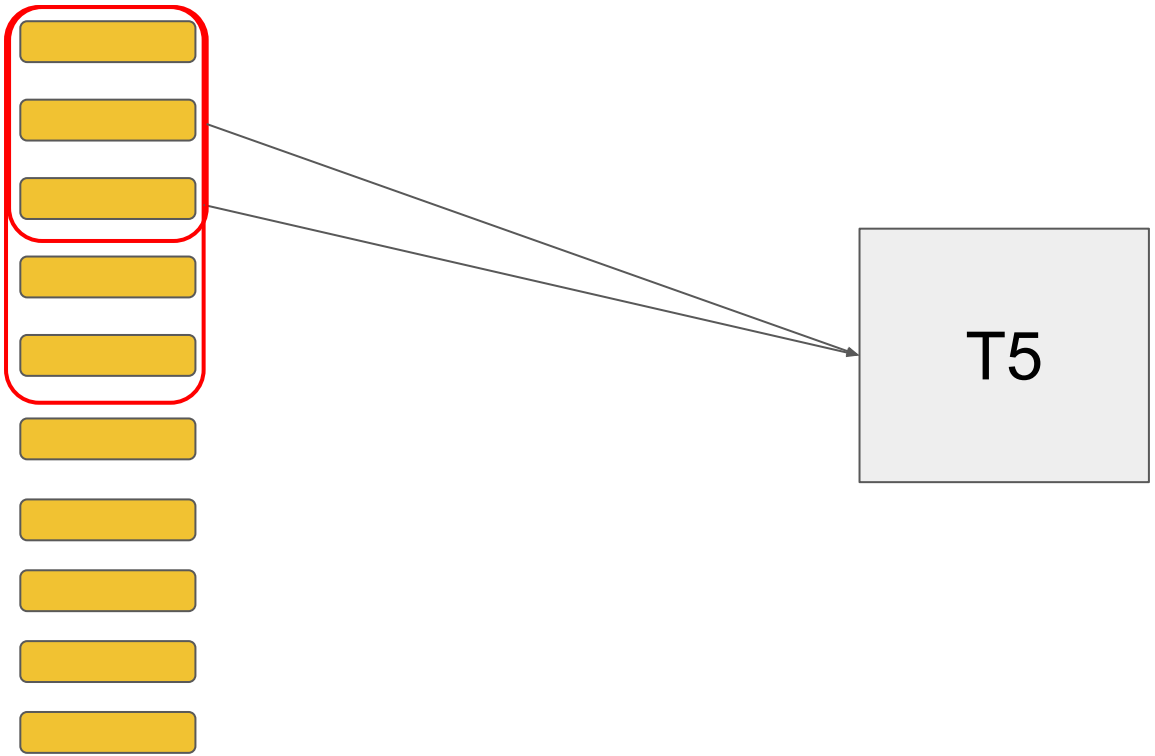


Knowledge Selection

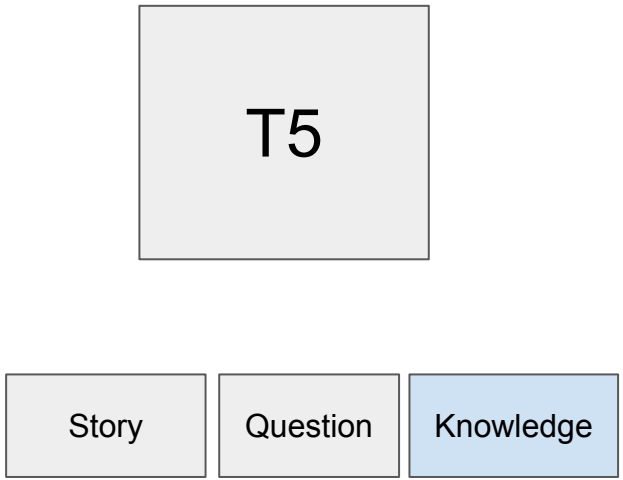
- Re-ranker + diversity metric



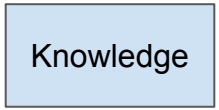
Amount of Knowledge



Knowledge Format



Knowledge Format

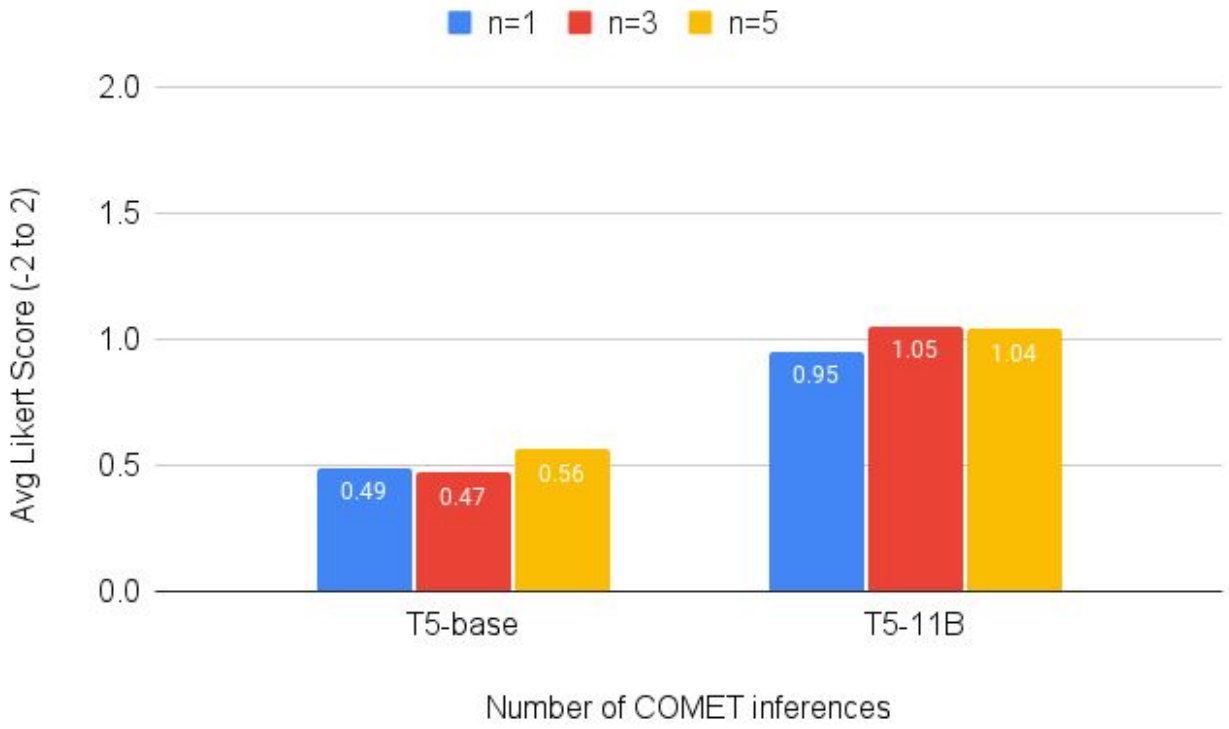


Tuple	relation: HasLastSubevent \n phrase: PersonX was in a deep sleep. \n
Verbalize	Sharon slept right through her alarm clock ends with she was in a deep sleep

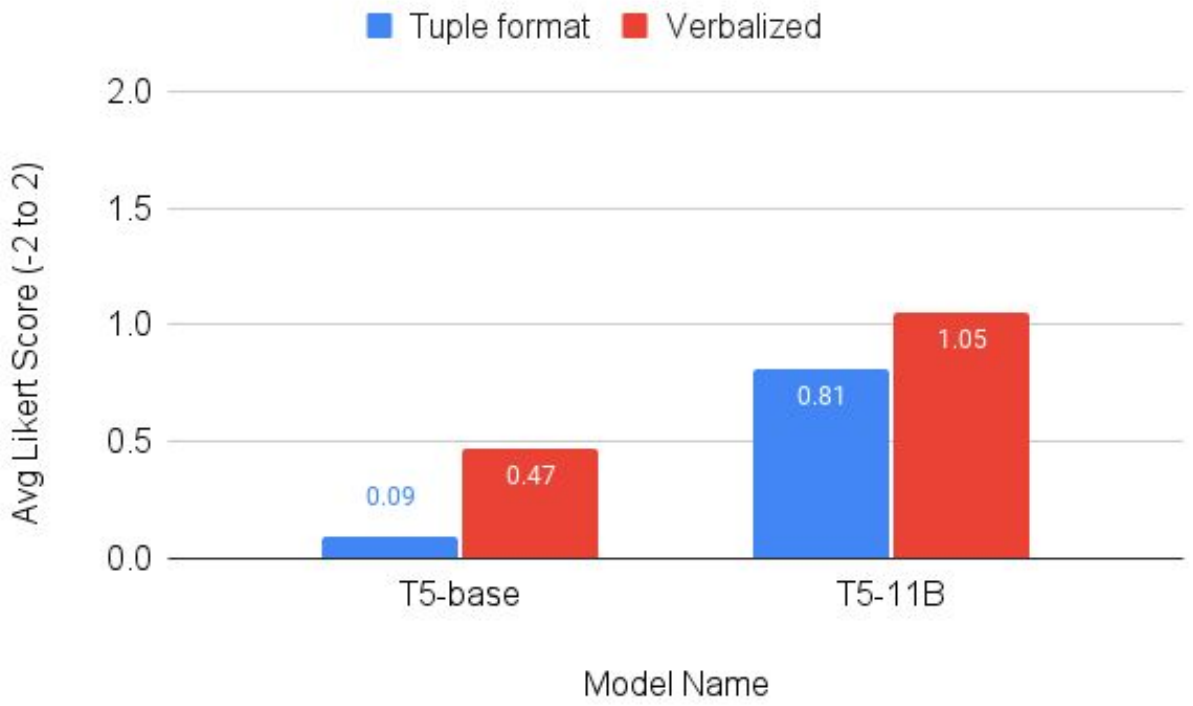
How do these Affect Models?

- Effect of Amounts of Knowledge
- Effect of Different Ways to Format Knowledge
- Effect of Knowledge Selection

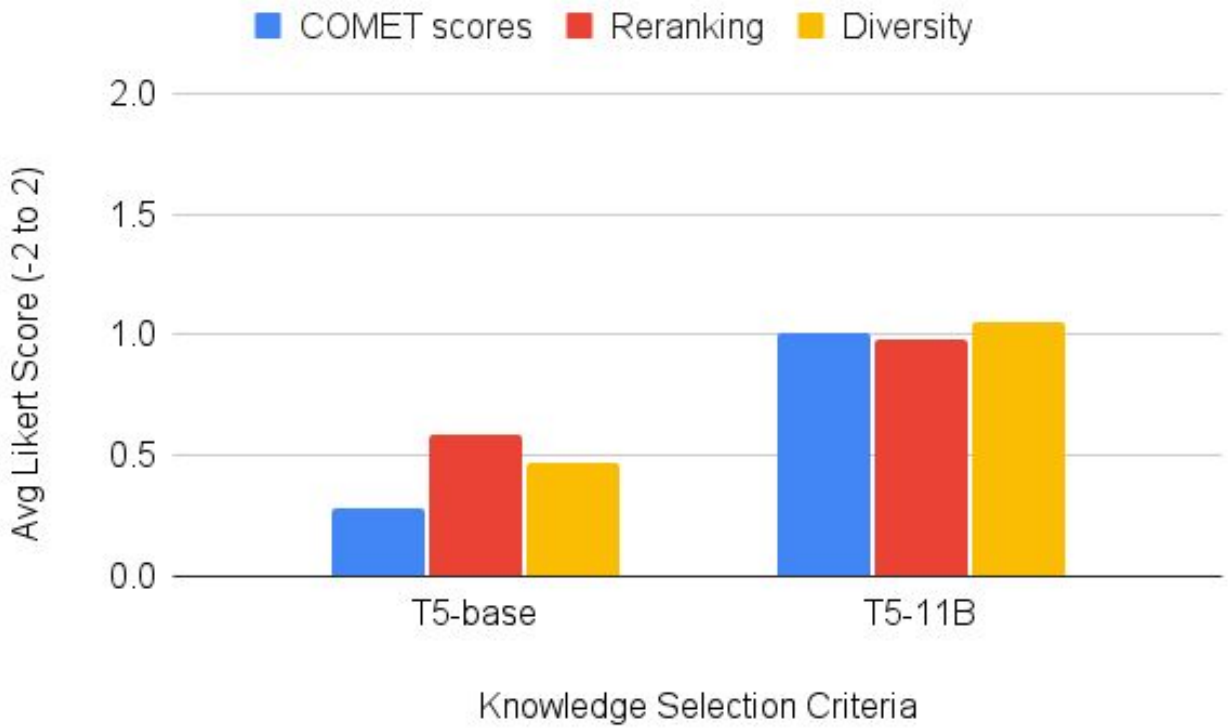
Effects of Amounts of Knowledge



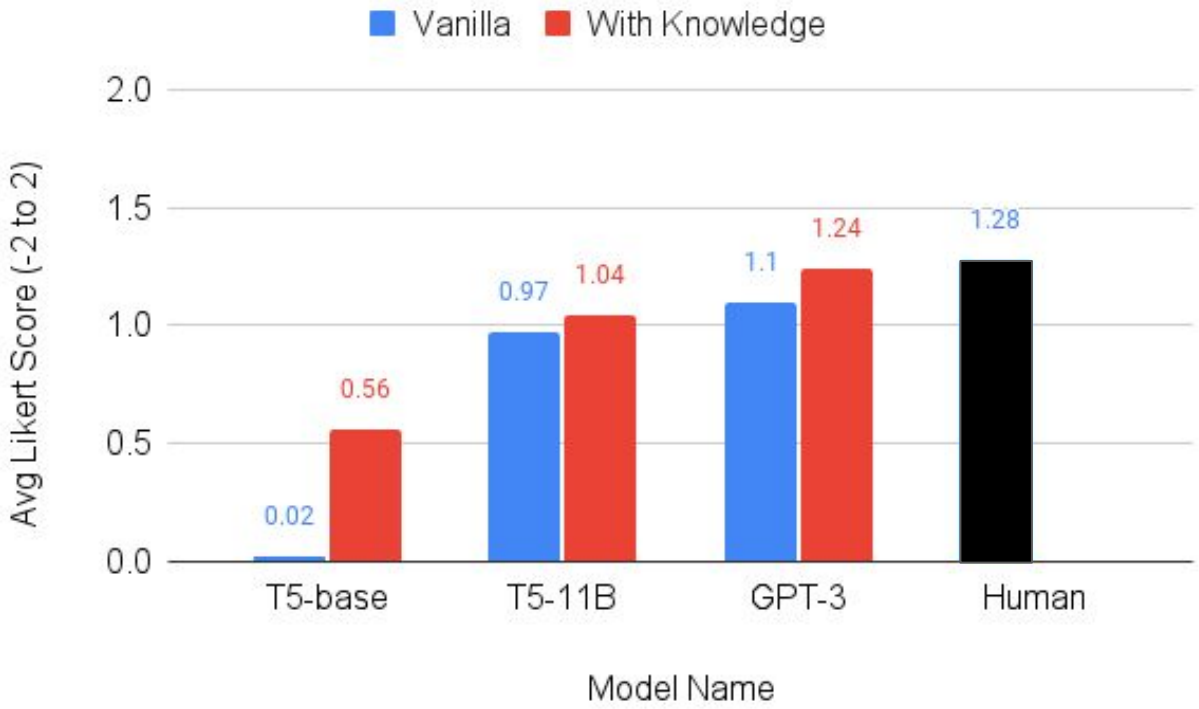
Effects of Different Ways to Format Knowledge



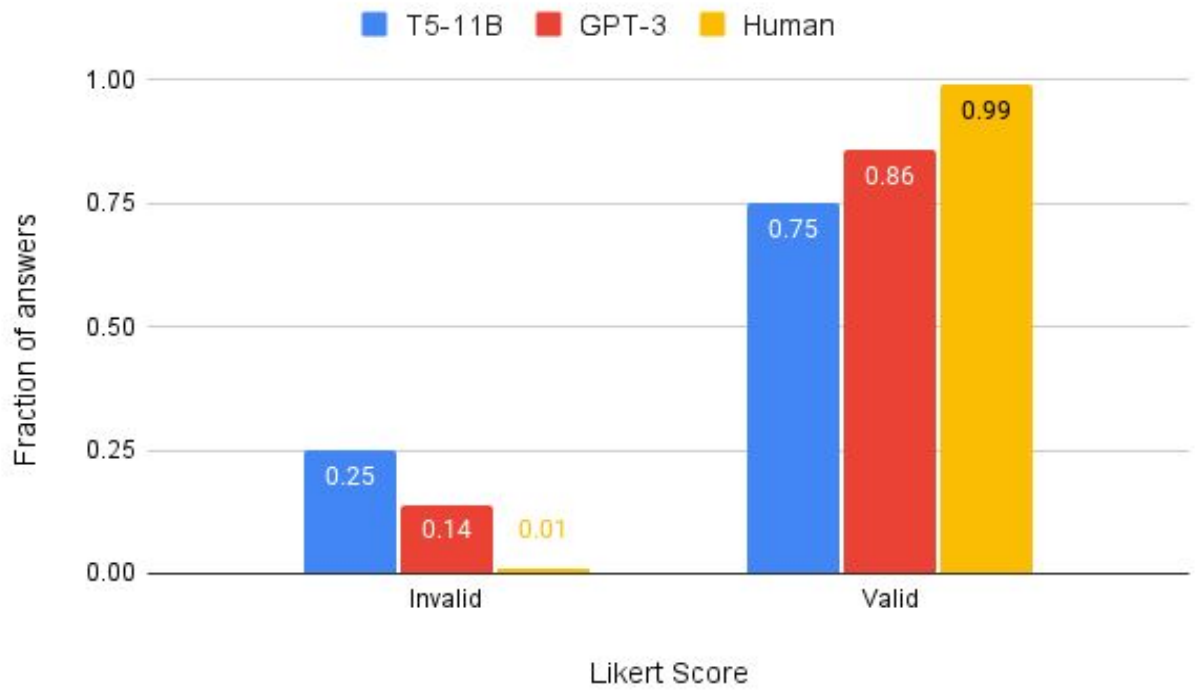
Effects of Knowledge Selection



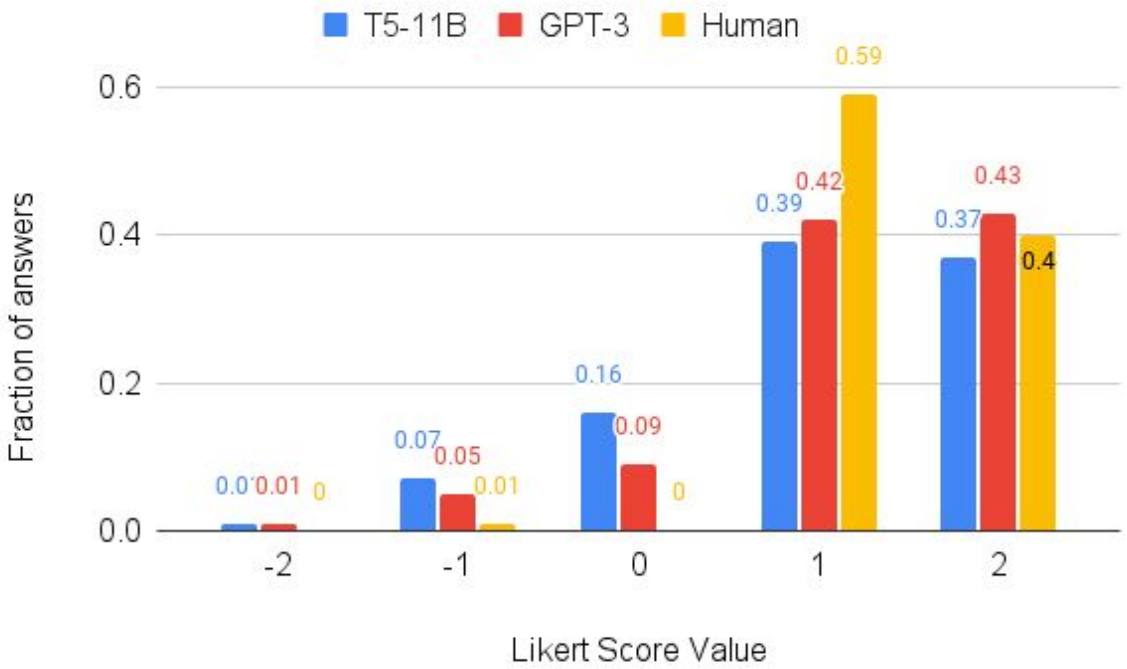
So, How Good are Models Now?



So, How Good are Models Now?



So, How Good are Models Now?



So, How Good are Models Now?

Matt and Sarah were pregnant. They wanted to announce it in a fun way. They wrote it on a cake. They invited their friends over. When their friends saw the cake, they were excited.

Q: Why did they write it?

To let their friends know that they were expecting a baby

+2

Matt and Sarah wanted to surprise their friends with something unexpected

+1

So, How Good are Models Now?

Maggie was drinking some green juice. She left the cup out awhile. When she went to get another sip it tasted odd. She realized that it had separated weirdly. She threw the juice out.

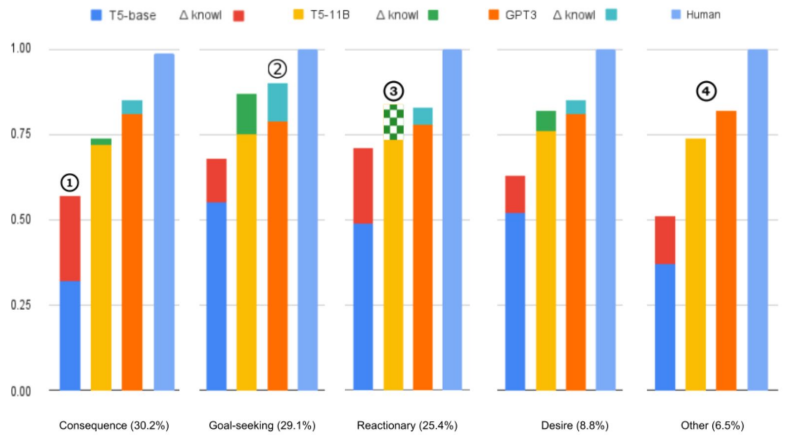
Q: Why did she leave the cup?

Maggie left the cup because it was too heavy | -2

Maggie had something else she had to attend to | +1

So, How Good are Models Now?

- Getting close on Average Likert scores
- Binary accuracy shows humans are significantly better
- Models come up with more obvious answers, but also get things wrong



Conclusions

- LLMs, if large enough, answer why questions fairly well
- Using diverse, ranked COMET inferences improves models of all sizes
- Models often produce more convincing answers, but humans are more consistent

Contact: ylal@cs.stonybrook.edu