

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



...

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.

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

TellMeWhy: A Dataset for Answering Why-Questions in Narratives

1 1 1 1 1
dataset, annotations & metadata

[View on GitHub](#) [Download in JSON format](#) [Download in CSV format](#)

TellMeWhy: A Dataset for Answering Why-Questions in Narratives

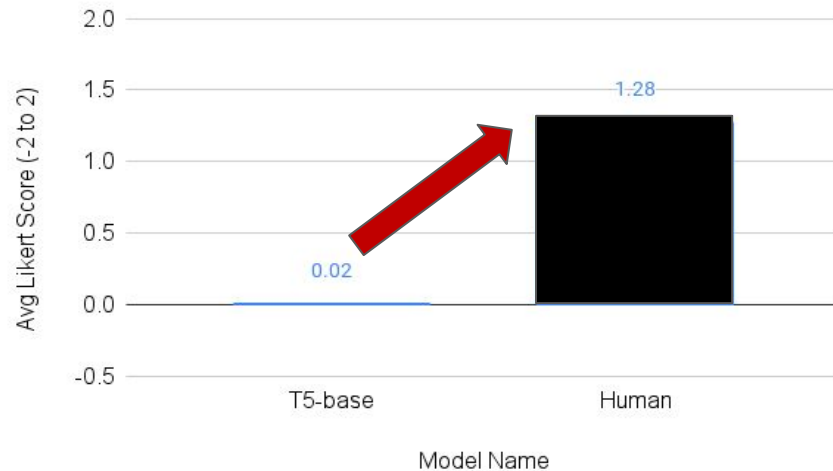
TellMeWhy is a large-scale crowdsourced dataset made up of more than **30k questions** and **free-form answers** concerning why characters in short narratives perform the actions described. Since a question can have many valid answers, we also released an easy-to-use **human evaluation** suite that should be used to correctly evaluate models for this why question answering task. Our paper "TellMeWhy: A Dataset for Answering Why-Questions in Narratives" published in Findings of ACL (ICNLP 2021). The camera-ready version is available on ArXiv here. The arXiv version is available here. It can also be found here. The video for the ACL Findings talk can be found here and the slides are here. This work was also presented in a poster session at the GEM workshop at ACL (ICNLP 2021).

Story: Sandra got a job at the zoo. She loved coming to work and seeing all of the animals. Sandra went to look at the polar bears during her lunch break. She watched them eat fish and jump in and out of the water. She took pictures and shared them with her friends.
Question: Why did Sandra go to look at the polar bears during her lunch break?
Answer: She wanted to take some pictures of them.

Dataset Information

Split	# Stories	# Questions
Train	2,588	23,964
Val	944	2,392
Test	944	3,090
Annotated Test	190	456
Total	9,636	30,519

[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?



T5

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?



T5



Matt and Sarah were pregnant



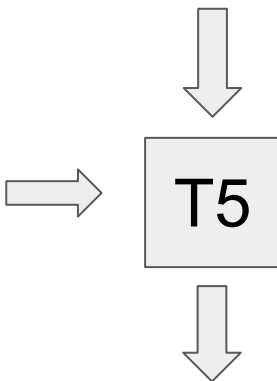
How Can We Improve Why QA Performance?

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?

T5



T5

How Can We Improve Access to Commonsense Knowledge

External Knowledge?

T5

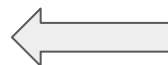


Commonsense
Knowledge
Resource

How Can We Improve Access to Commonsense Knowledge

External Knowledge?

T5



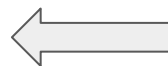
GLUCOSE Dataset

COMET 

How Can We Improve Access to Commonsense Knowledge

External Knowledge?

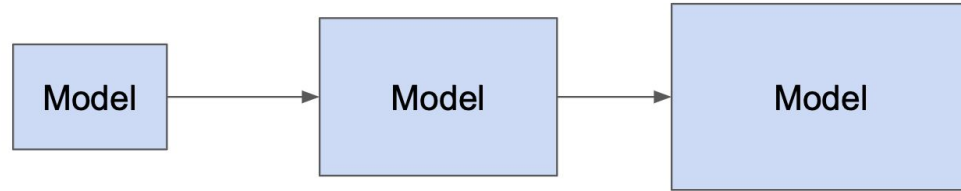
T5



COMET 

Our Contributions

Larger Models

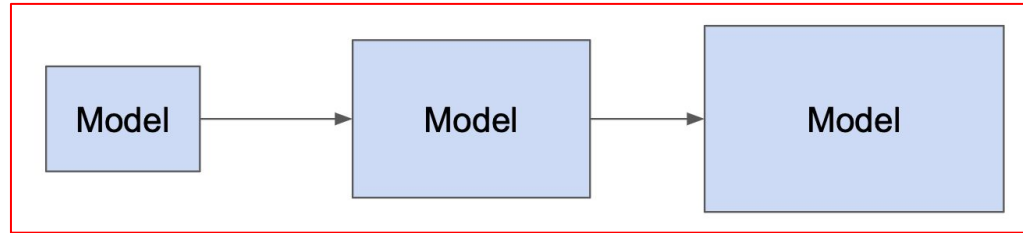


External Knowledge



Our Contributions

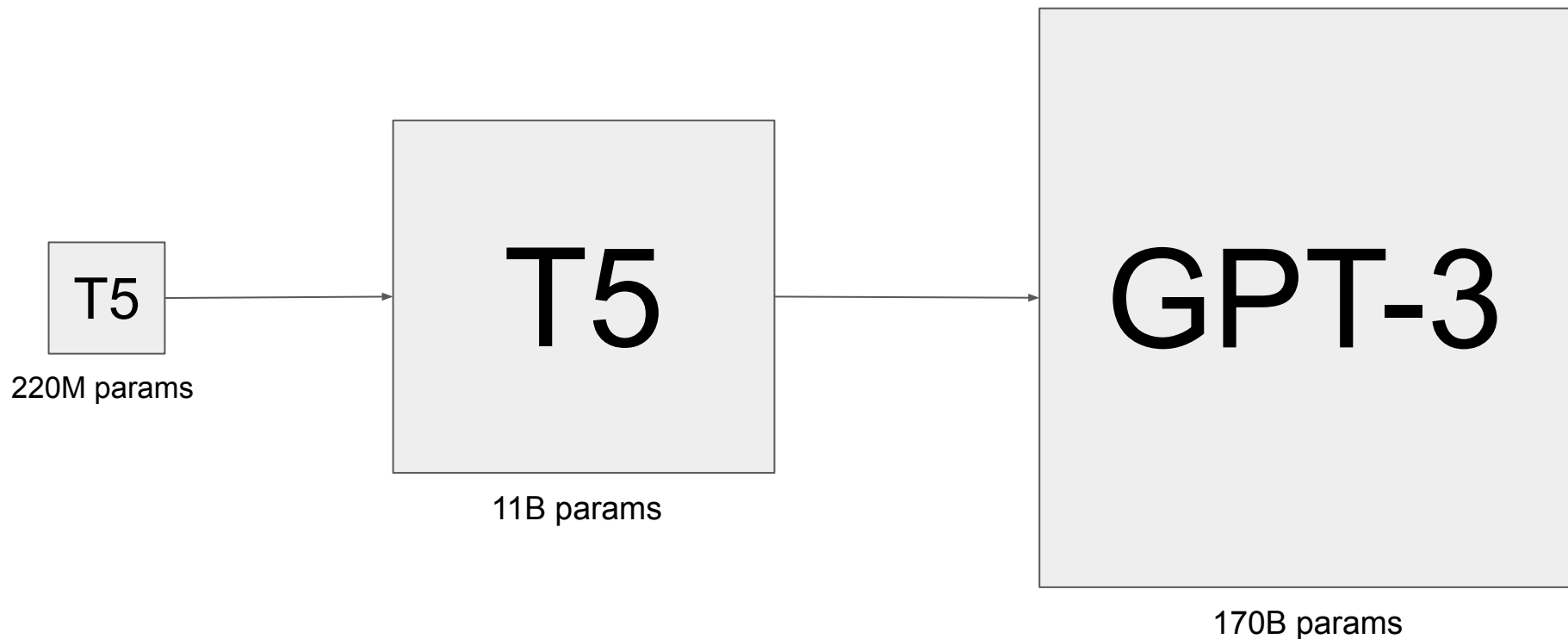
Larger Models



External Knowledge



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)

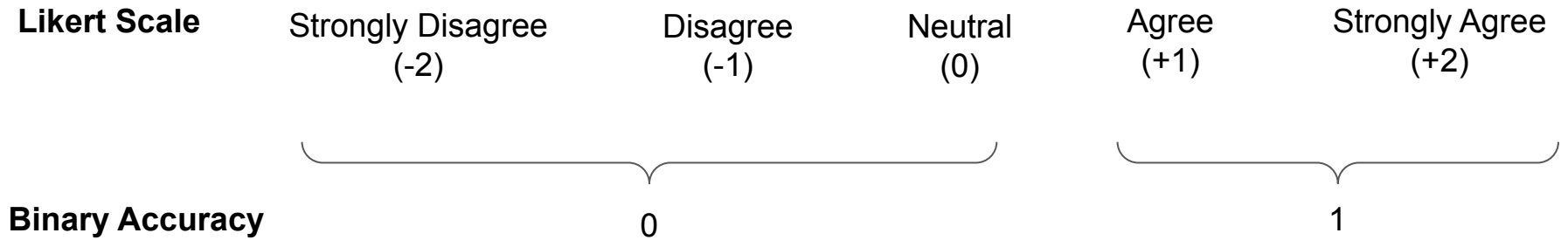
Agree
(+1)

Strongly Agree
(+2)

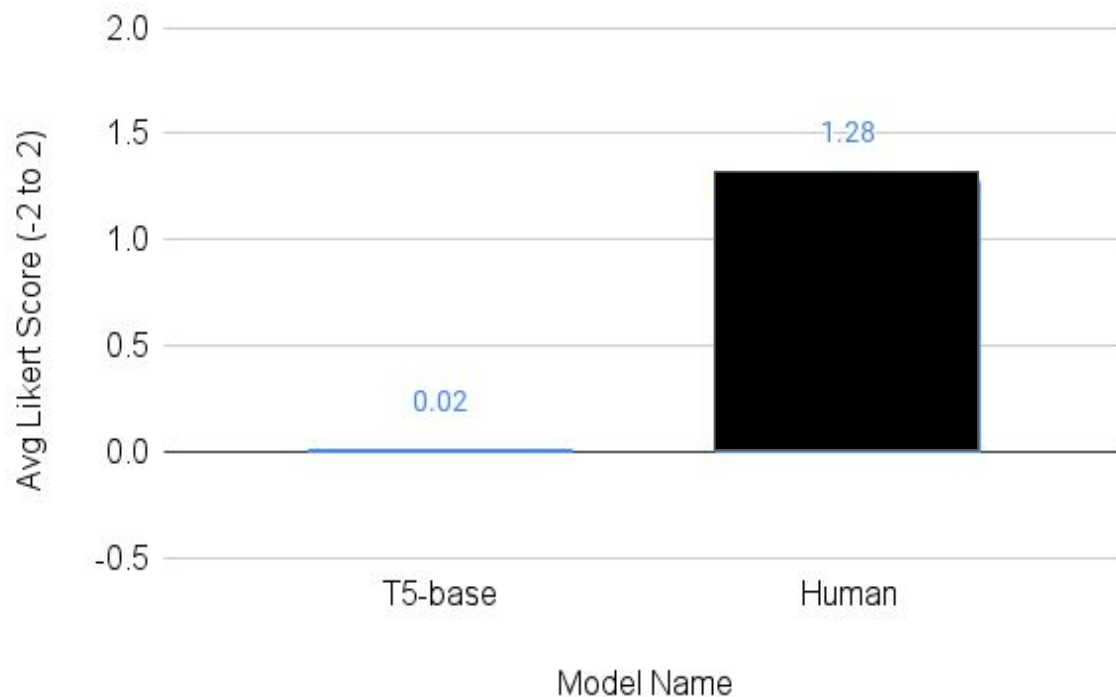
Average Likert Score



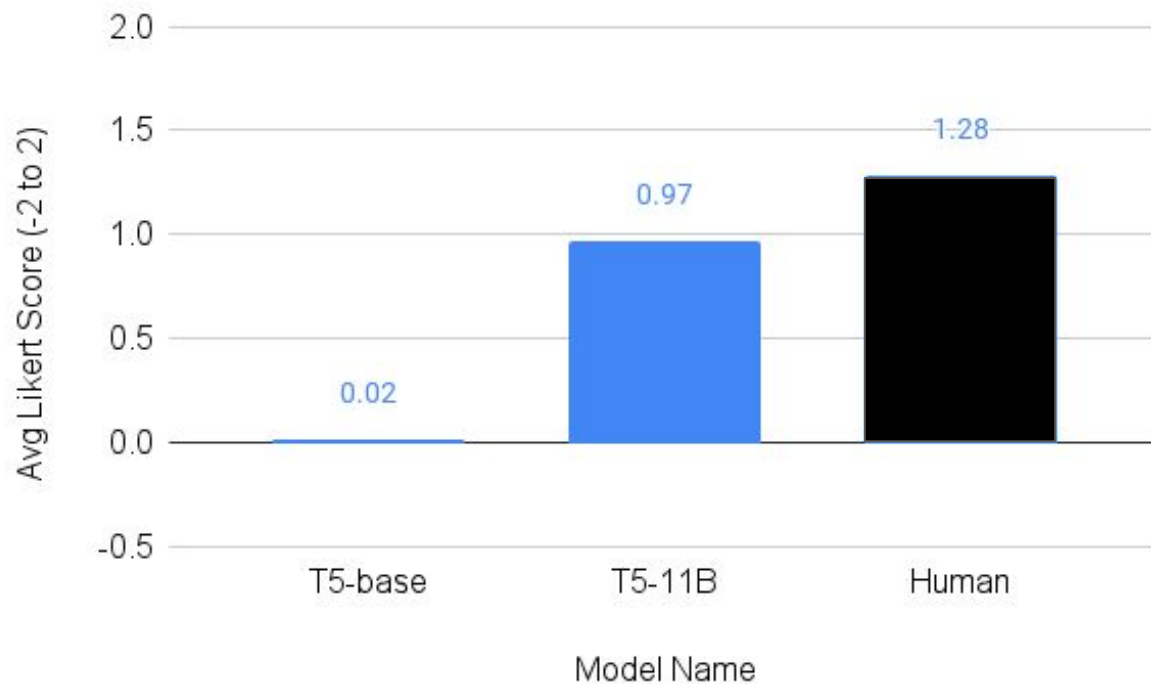
Evaluation Setup



Using Larger Models

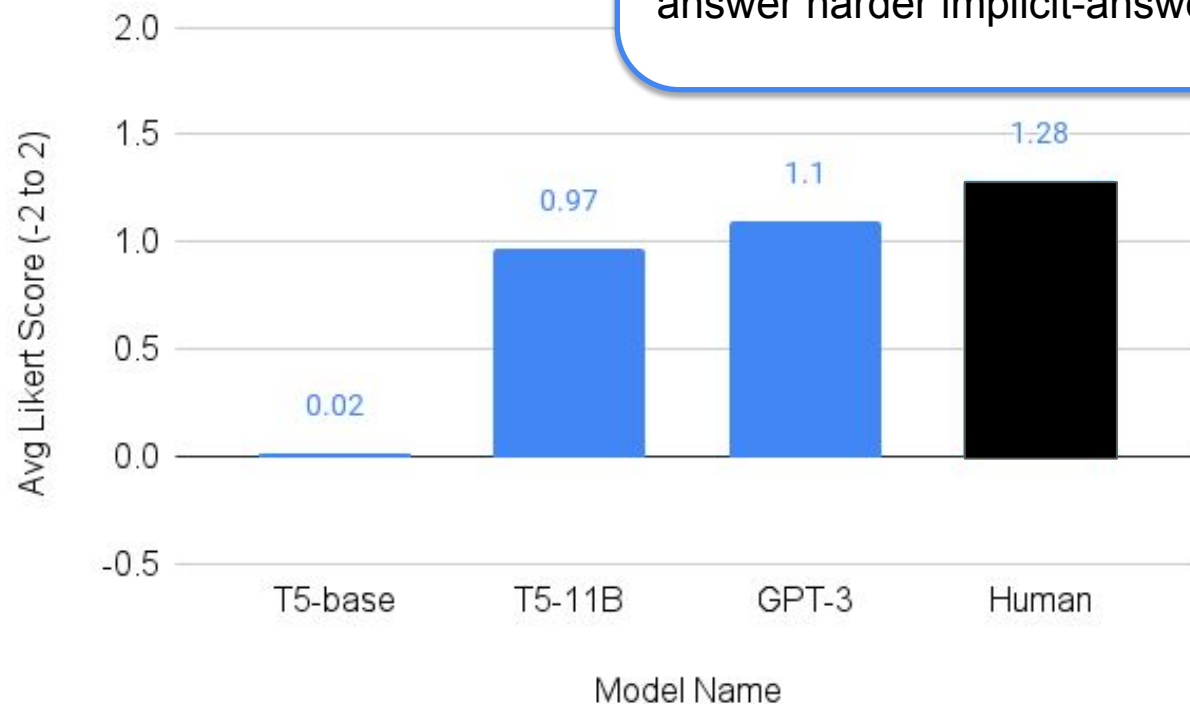


Using Larger Models



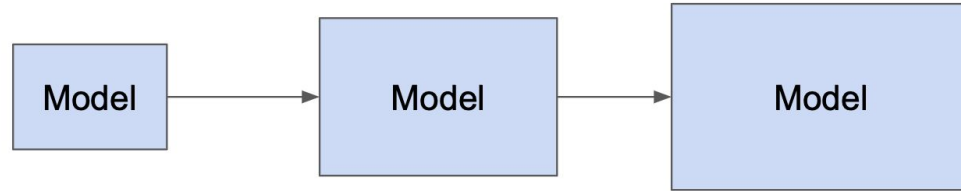
Using Larger Models

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

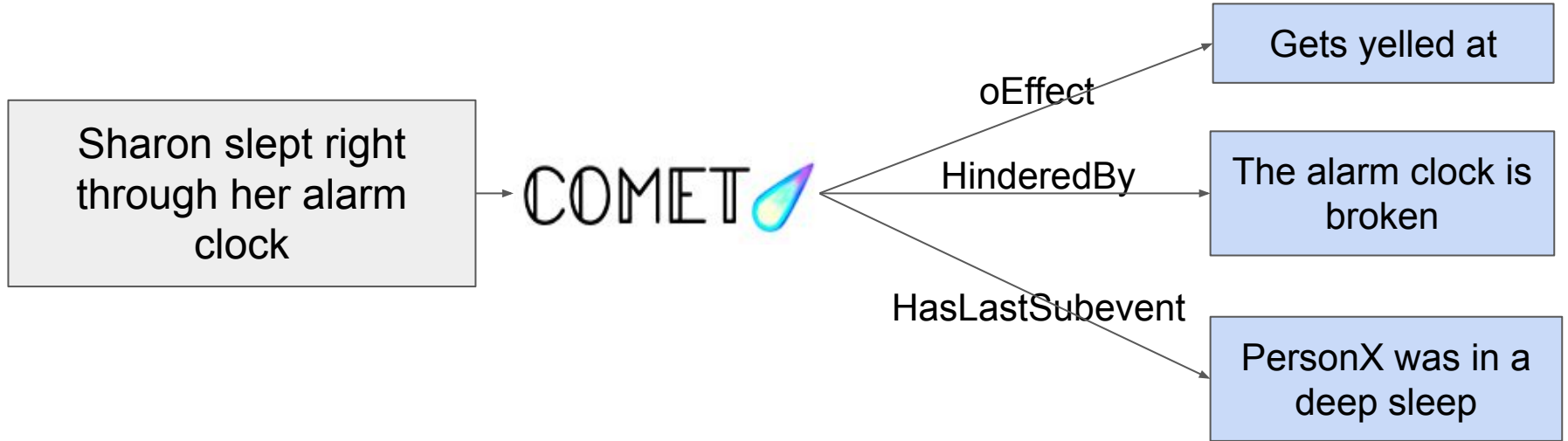
Sharon slept right through her alarm clock! She was dreaming of the ocean and how the waves sounded. She didn't even hear her family leave for church. When she woke up and everyone was gone, she was afraid. She couldn't believe they had gone without her.

COMET

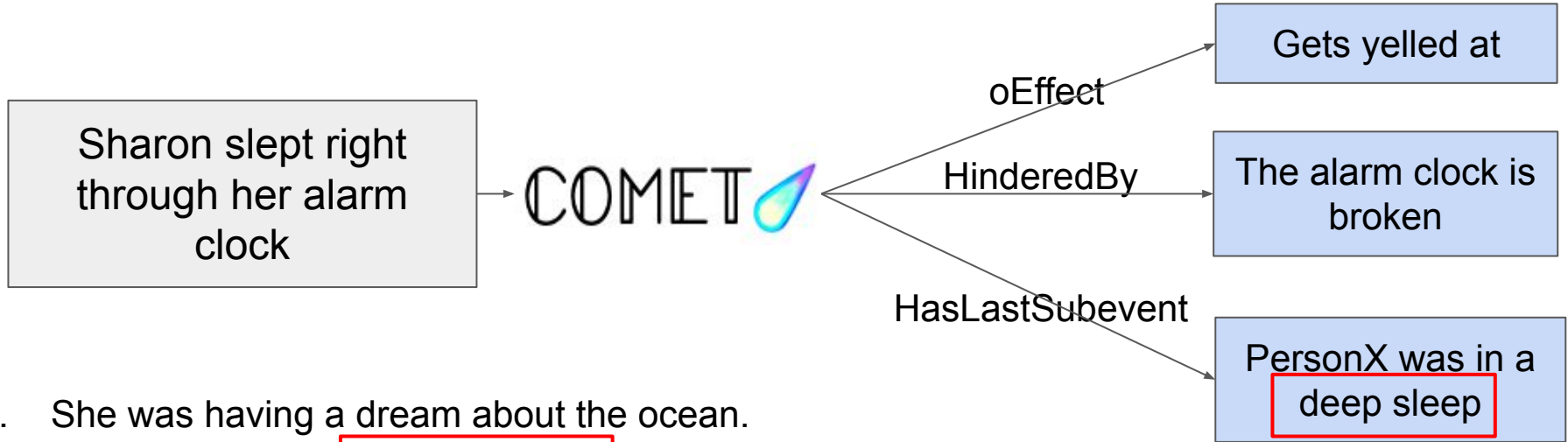
Sharon slept right through her alarm clock! She was dreaming of the ocean and how the waves sounded. She didn't even hear her family leave for church. When she woke up and everyone was gone, she was afraid. She couldn't believe they had gone without her.

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

COMET



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

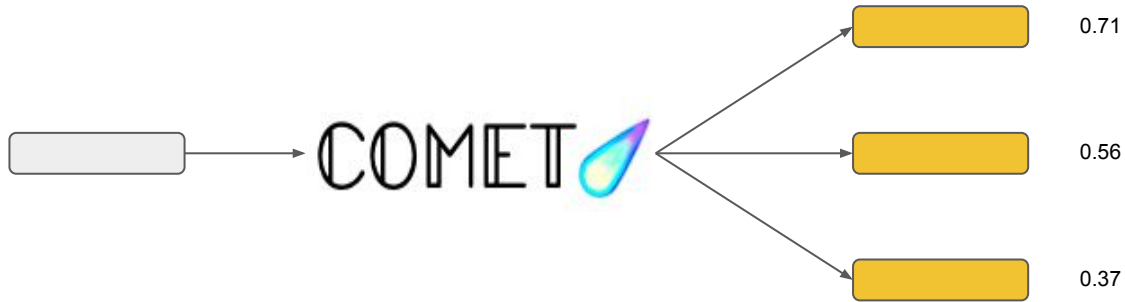
Models are known to be sensitive to input structure

Knowledge Selection

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

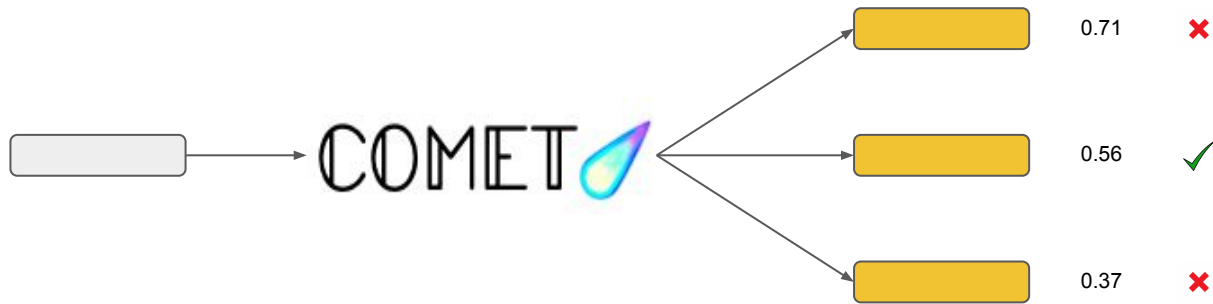
Knowledge Selection

- COMET scores



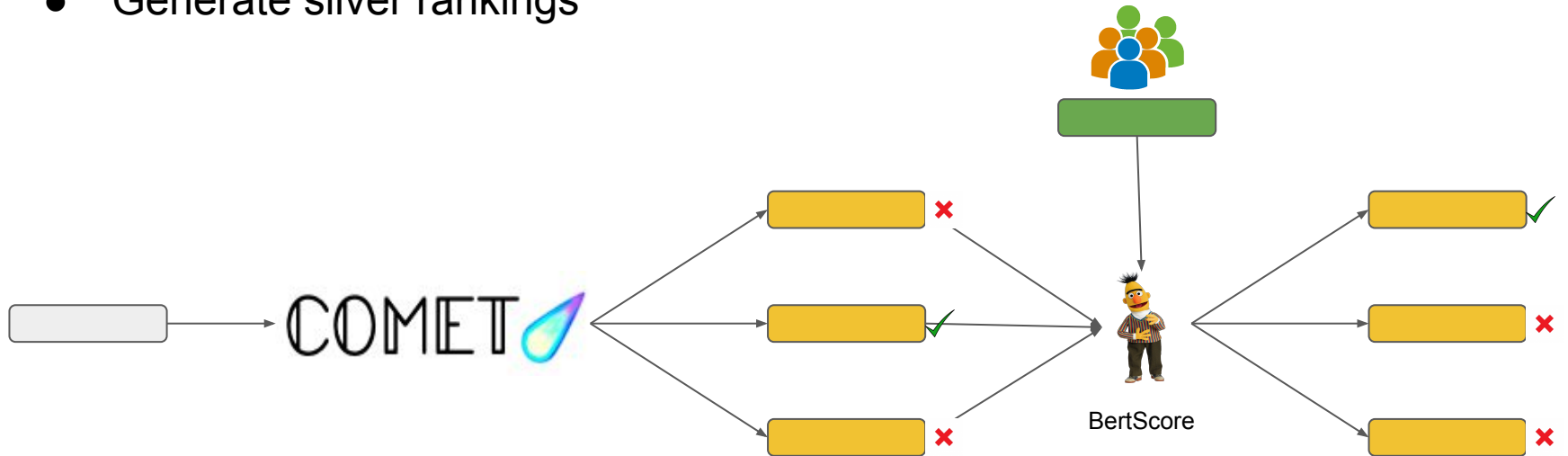
Knowledge Selection

- Trained re-ranking model



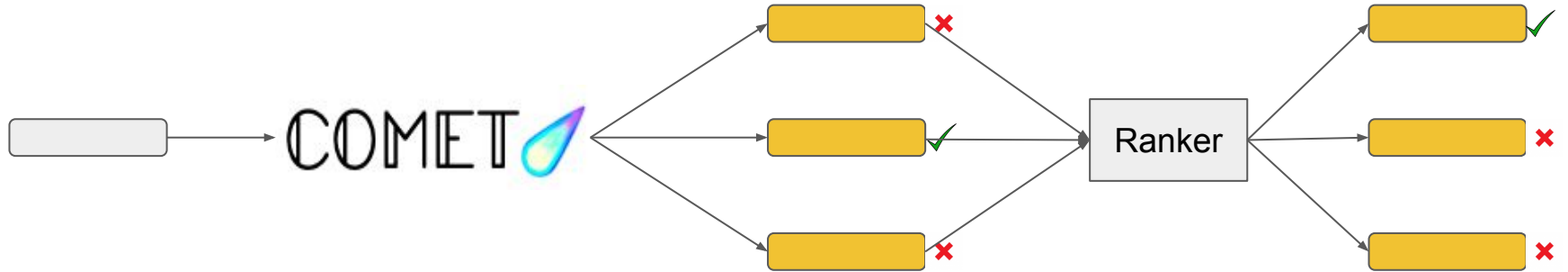
Knowledge Selection

- Generate silver rankings



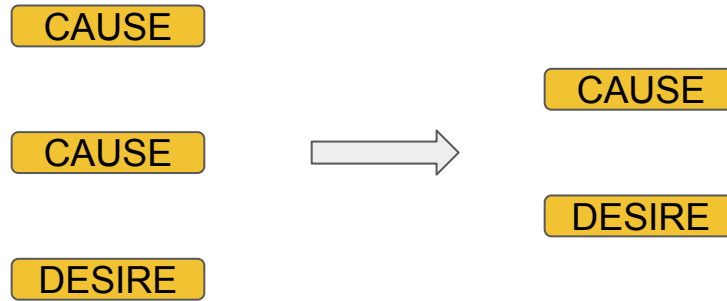
Knowledge Selection

- Generate silver rankings
- Train a re-ranker

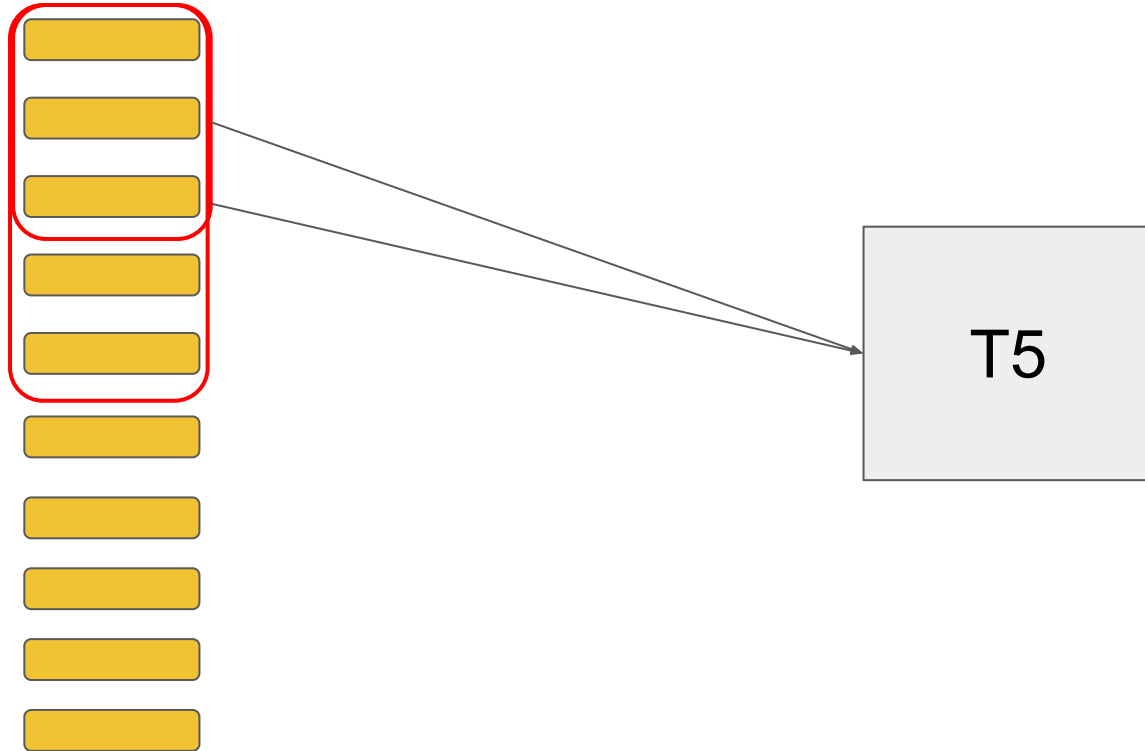


Knowledge Selection

- Re-ranker + diversity metric



Amount of Knowledge



Knowledge Format

T5



Knowledge Format

Knowledge

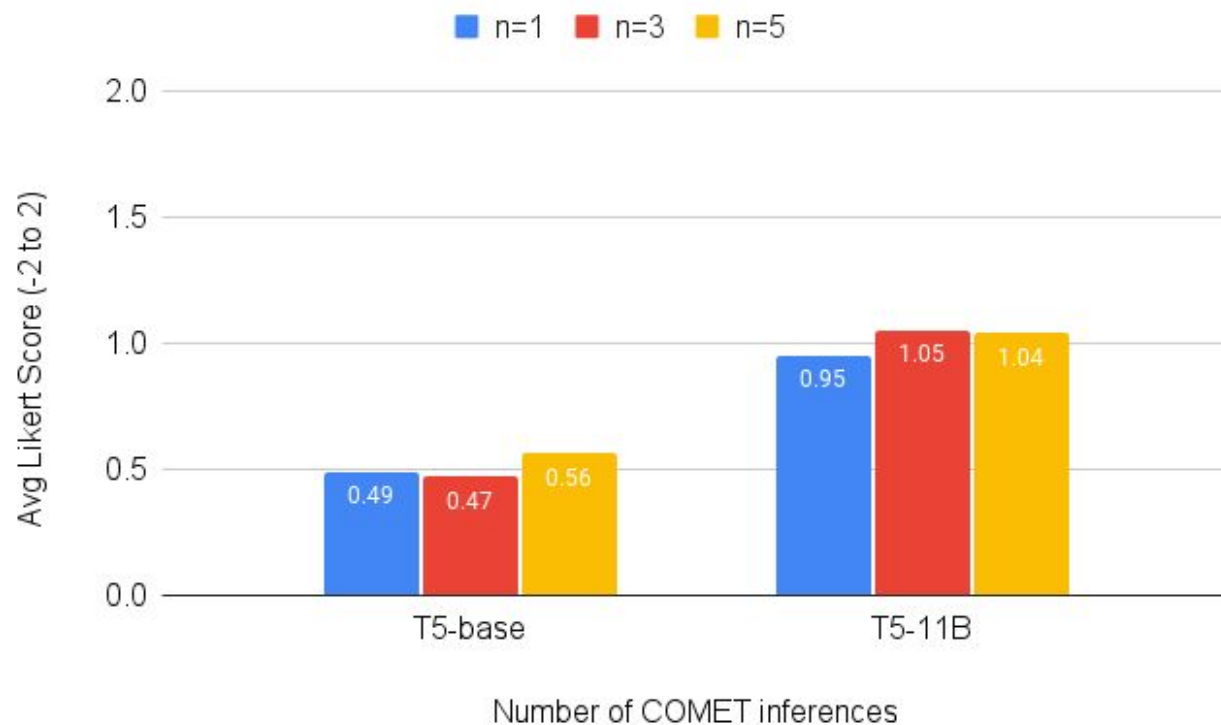
Tuple relation: **HasLastSubevent** \n phrase: PersonX was in a deep sleep. \n

Verbalize Sharon slept right through her alarm clock **ends with** she was in a deep sleep

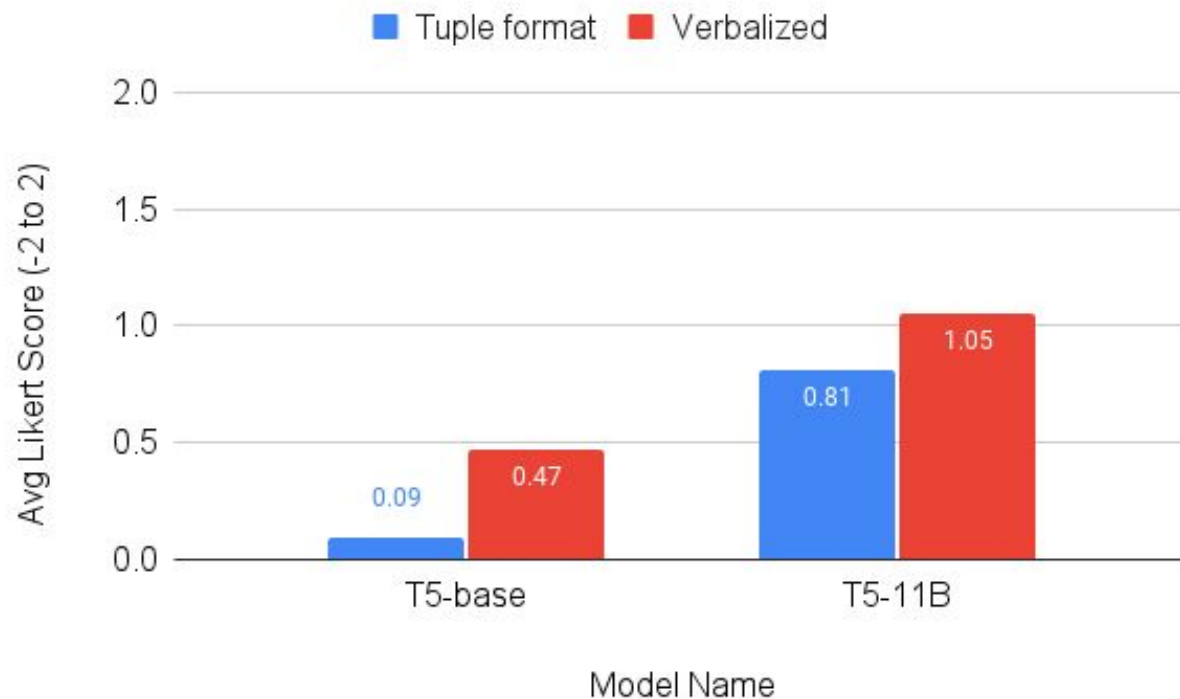
How do these Affect Models?

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

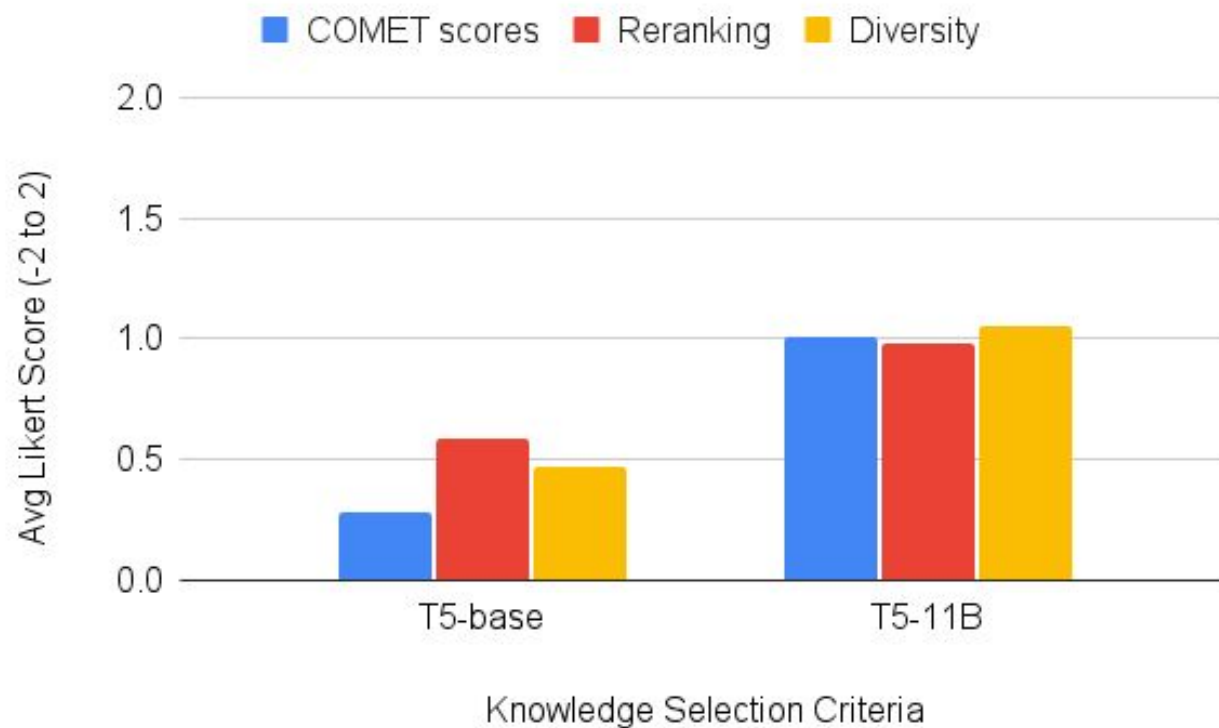
Effects of Amounts of Knowledge



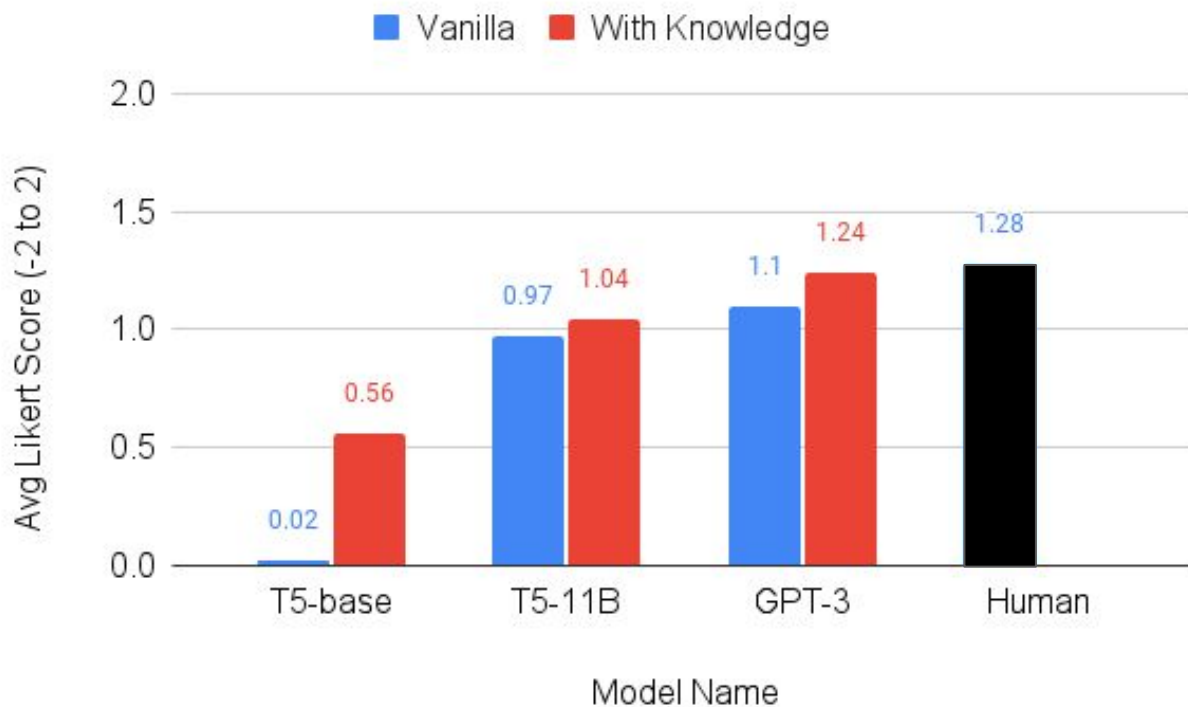
Effects of Different Ways to Format Knowledge



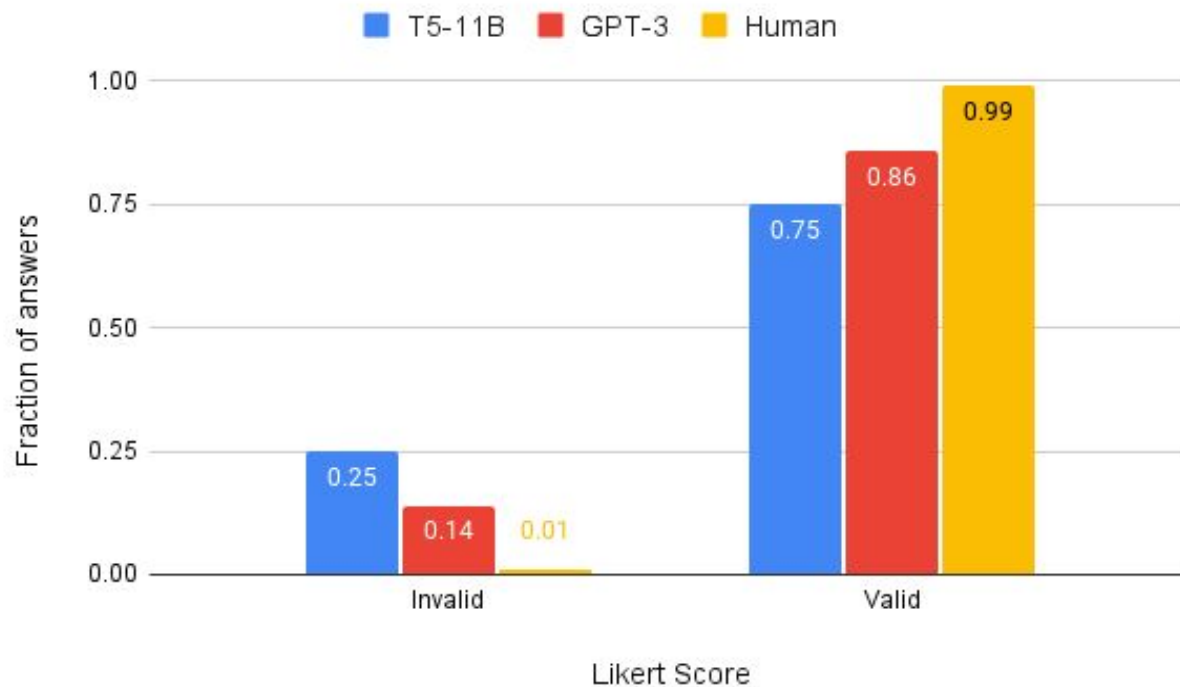
Effects of Knowledge Selection



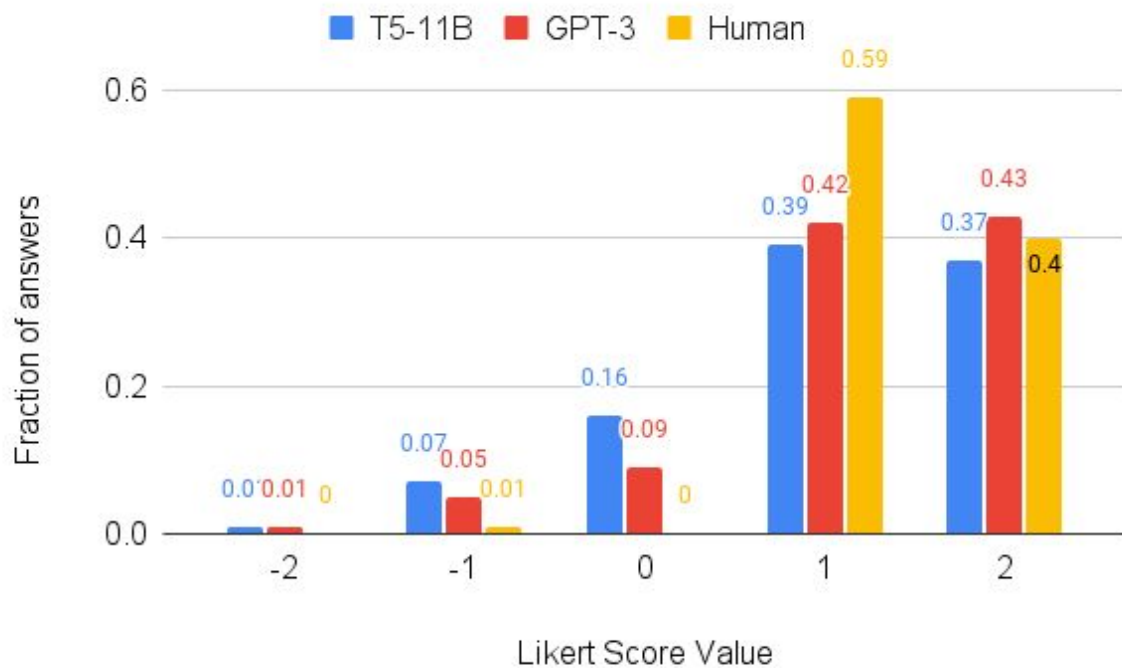
So, How Good are Models Now?



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

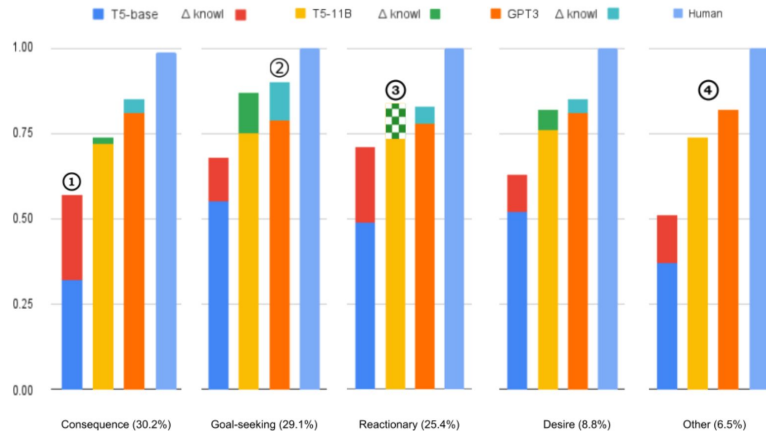


Maggie had something else she had to attend to

+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