



# TellMeWhy: A Dataset for Answering Why-Questions in Narratives



Yash Kumar Lal



Nate Chambers



Raymond Mooney



Niranjan  
Balasubramanian



# Knowing *why* is important for reasoning about events.

**Story:** Rudy was convinced that bottled waters all tasted the same. **He went to the store and bought several popular brands.** He went back home and set them all on a table. He spent several hours tasting them one by one. He came to the conclusion that they actually did taste different.

**Q:** Why did Rudy go to the store and buy several popular brands?

**A:** He wanted to taste test.

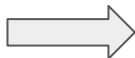
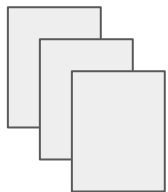
Knowing why can help

- Explain his motivation
- Visualize events in a narrative (Kintsch & Dijk, 1978)
- Understand plans and goals (Schank & Abelson, 1975)

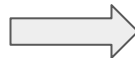
No large scale datasets for ***why*** questions **over narratives**

# TellMeWhy: A collection of why questions over narratives

ROCStories



**Story:** Rudy was convinced that bottled waters all tasted the same. **He went to the store and bought several popular brands.** He went back home and set them all on a table. He spent several hours tasting them one by one. He came to the conclusion that they actually did taste different.



**Q:** Why did Rudy go to the store and buy several popular brands?

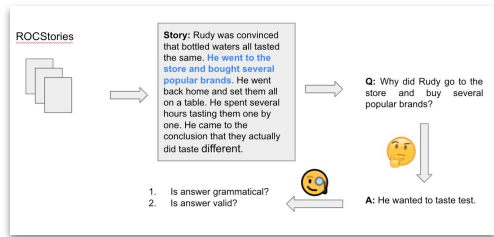


**A:** He wanted to taste test.



1. Is answer grammatical?
2. Is answer valid?

# TellMeWhy: Dataset Characteristics



Implicit Answers

29%

Explicit Answers

71%

☐ Answers not always in text.

Token Overlap  
b/w Turker Answers

26%

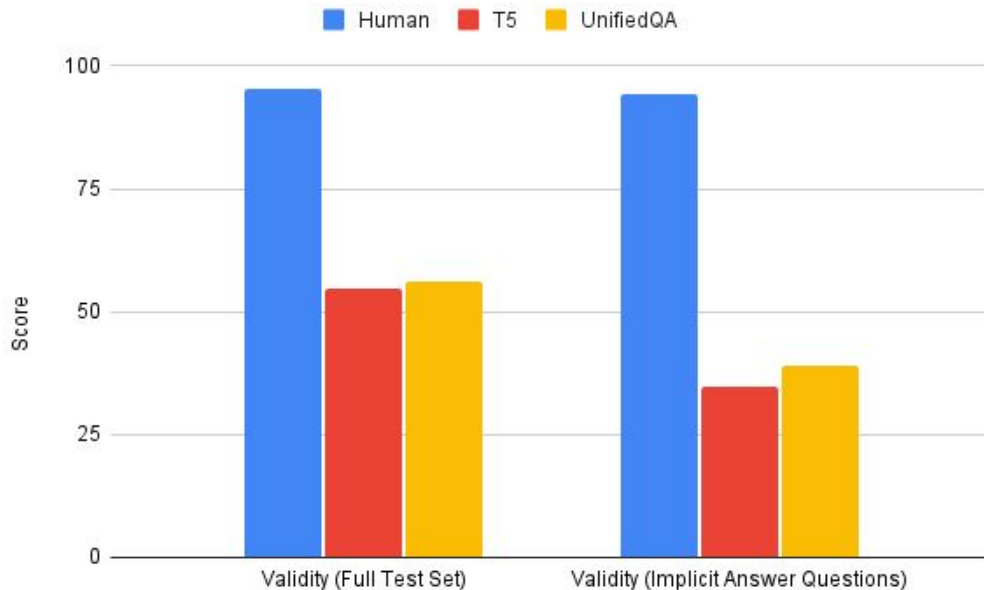
☐ Diverse answers possible.

# Benchmarking Large LM-based QA Models on TellMeWhy.

## Models

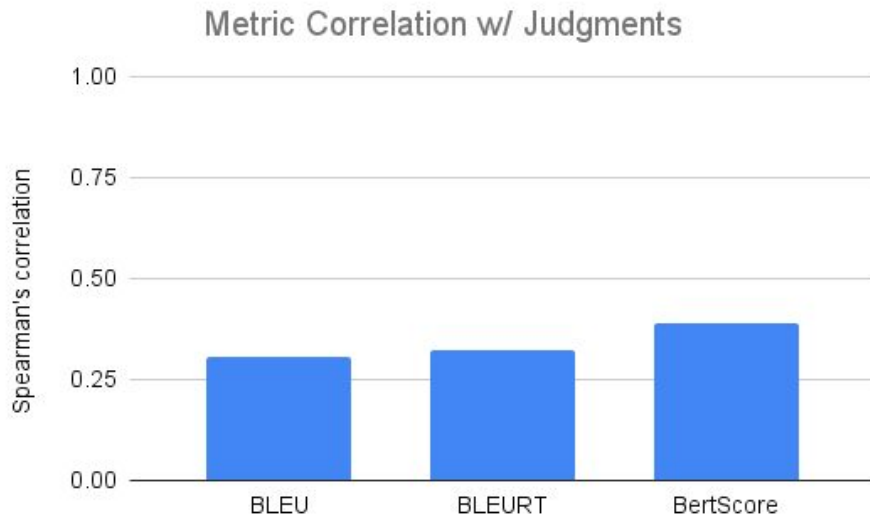
T5 (Raffel et al, 2020)

UnifiedQA (Khashabi et al, 2020)



Fine-tuning large QA models isn't enough.

# Automatic evaluation is inadequate.



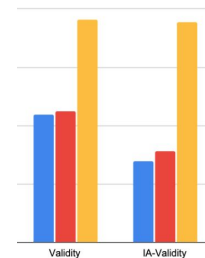
- ❑ Automatic metrics not enough (unsurprisingly).
- ❑ BertScore is the most reliable (relatively).

# Takeaways

A challenging why-question dataset.

Large LMs are not capable of answering why questions.

Harness for standardized human evaluation for TellMeWhy.



**Answering questions based on a story**

[Show the instructions. Please click if this is your first time!](#)

Contact us:

**Task**

Story: Anna could not swim. She decided it was time to learn. She signed up for a class at the pool. She began by learning slow, easy strokes. Soon she was swimming quickly and eagerly.

Questions: Why did she decide it was time to learn?

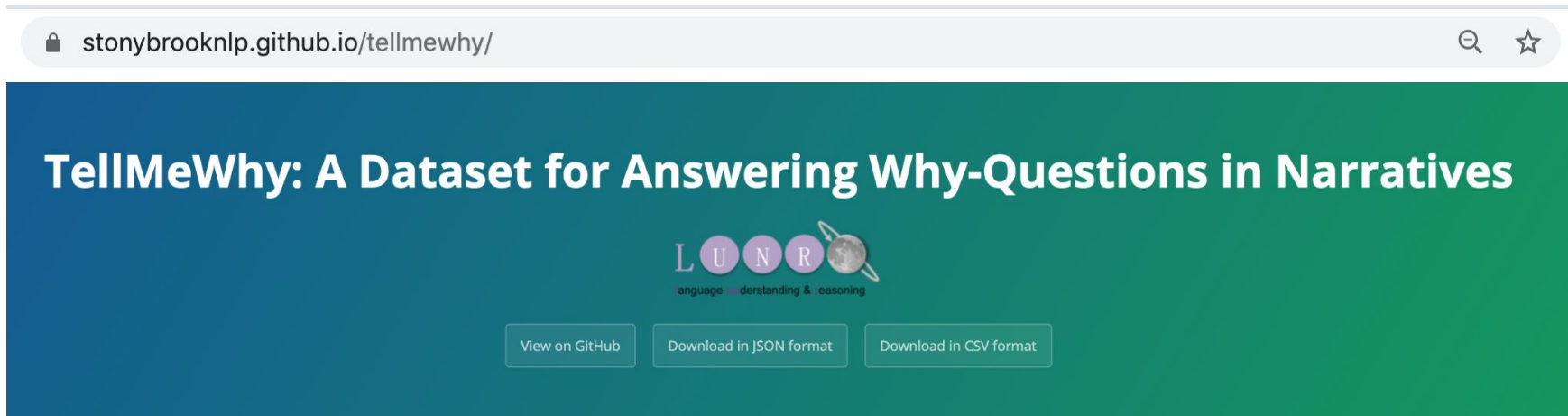
In these same information in the story that can help you answer this question? (Even if not, please still enter an answer below). Yes No Question not comprehensible

She decided it was time to learn because...

Please select the sentences that most helped you decide your answer.


- Anna could not swim.
- She decided it was time to learn.
- She signed up for a class at the pool.
- She began by learning slow, easy strokes.
- Soon she was swimming quickly and eagerly!

Code and data: <https://stonybrooknlp.github.io/tellmewhy/>



stonybrooknlp.github.io/tellmewhy/

# TellMeWhy: A Dataset for Answering Why-Questions in Narratives

L U N R   
language understanding & reasoning

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

Contact: [ylal@cs.stonybrook.edu](mailto:ylal@cs.stonybrook.edu)