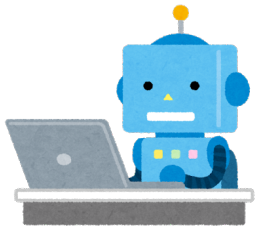
ACL 2022 #1566

What Makes Reading Comprehension Questions Difficult?





Saku Sugawara,[1] Nikita Nangia,[2] Alex Warstadt,[2] Samuel R. Bowman[2]

[1]National Institute of Informatics, [2] New York University

saku@nii.ac.jp

NII National Institute of Informatics

NYU

ML^2 Machine Learning for Language

Our Goal

We want large-scale NLU benchmark data that is:

- **Difficult enough** to **discriminate** state of the art models
- **Linguistically diverse** to **validate** good performance

This is a part of the criteria for better NLU benchmarking in Bowman & Dahl (2021)

Background: Crowdsourcing NLU Data

- Protocols of worker handling and feedback (Nangia et al., 2021)
- Design of the collection task (Ning et al., 2020; Rogers et al., 2020)

What aspects of **text sources** affect the difficulty and diversity of examples?

Motivation

MCTest: Tony walked home from school on his birthday. He was surprised to see a lot of cars in front of his house. When he opened the door and entered the house, he heard a lot of people yell, "Surprise!" It was a surprise party for his birthday. His parents called all his friends' parents and invited them to come to a party for Tony. [...]
Q: *Who were invited to the party and by who?*
☐ *Tony's parents invited only his friends*
☐ *Tony invited his friends and their parents*
☐ *Tony's parents invited his friends' parents*
☑ *Tony's parents invited his friends and their parents*

Children-level story
→ Easy to read
& factoid / simple math..?.

ReClor: Humanitarian considerations aside, sheer economics dictates that country X should institute, as country Y has done, a nationwide system of air and ground transportation for conveying seriously injured persons to specialized trauma centers. Timely access to the kind of medical care that only specialized centers can provide could save the lives of many people. [...]
Q: *What is the economic argument supporting the idea of a transportation system across the nation of Country X?*
☐ *Building the transportation system creates a substantial increase of jobs for the locals*
☑ *Increasing access to specialized medical centers can lower the chance of the workforce population dying*
☐ *Transportation ticket prices directly contribute to the government's revenue*
☐ *Country Y was successful with their attempts to potentially save lives so Country X should try it as well*

Technical document
→ Difficult to read
& logical reasoning...?

- The more difficult a passage to read, the more challenging a question about it?
- Are specific domains useful for collecting specific types of questions?

This Study: What Passage Sources are Useful?

Research Question

- Are difficult passages more suitable for crowdsourcing challenging questions?

Method

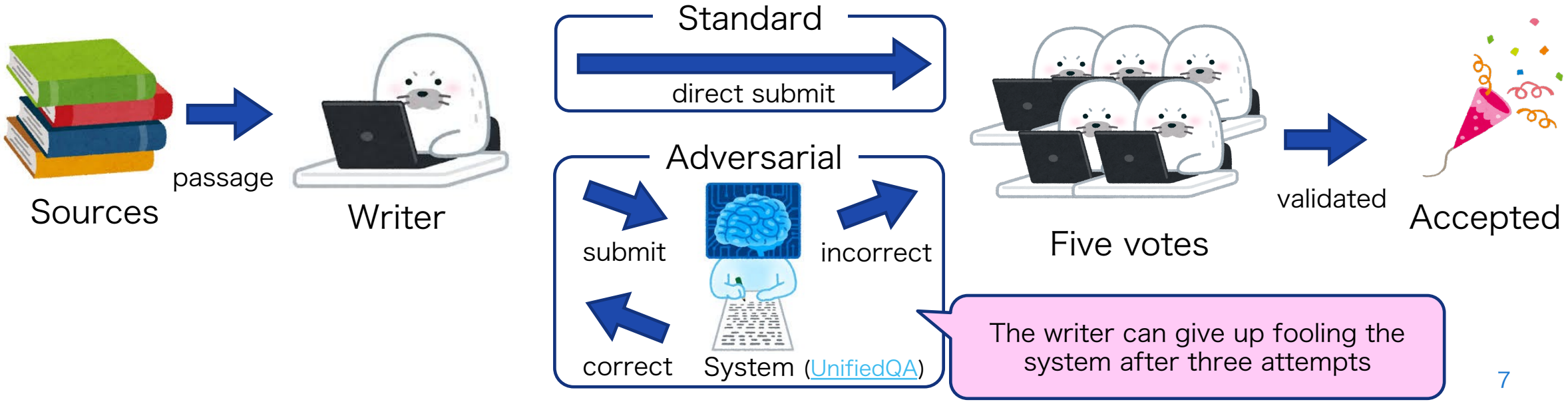
- Crowdsource reading comprehension questions using passages taken from seven different passage sources
- Analyze how question difficulty and type are affected by linguistic aspects of passages (e.g., source, readability, syntactic & lexical surprisal, and vocabulary)
- Bonus: what if we do the same data collection with model-in-the-loop?

Passage Sources

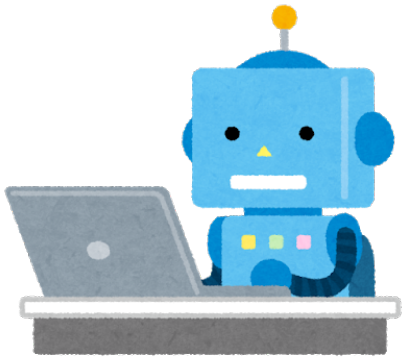
1. MCTest (children's stories Richardson et al., 2013)
2. Project Gutenberg (fictions: novel, narrative, story)
3. Slate (online magazine articles in Open ANC; Ide and Suderman, 2006)
4. RACE (middle- and high-school English exams; Lai et al., 2017)
5. ReClor (exercise questions for GMAT and LSAT exams; Yu et al., 2020)
6. Science articles in Wikipedia
7. Arts articles in Wikipedia

Crowsdourcing: Question Writing and Validation

- We ask crowdworkers to write a question with four options given a passage
- Workers are assigned to either the standard or adversarial data collection
- Five different workers validate each question



Statistics & Experiments



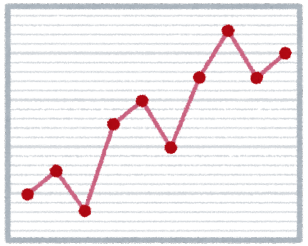
Dataset

- Initially collected: 4,340 questions
  - 620 Qs * 7 sources, 310 each for the standard and adversarial methods
  - Validated ≈ 90%
  - High-agreement ≈ 65%

Systems (8 models)

- RoBERTa large (fine-tuned on RACE) * 4 different models
- DeBERTa large & xlarge (fine-tuned on MNLI or not): 4 models

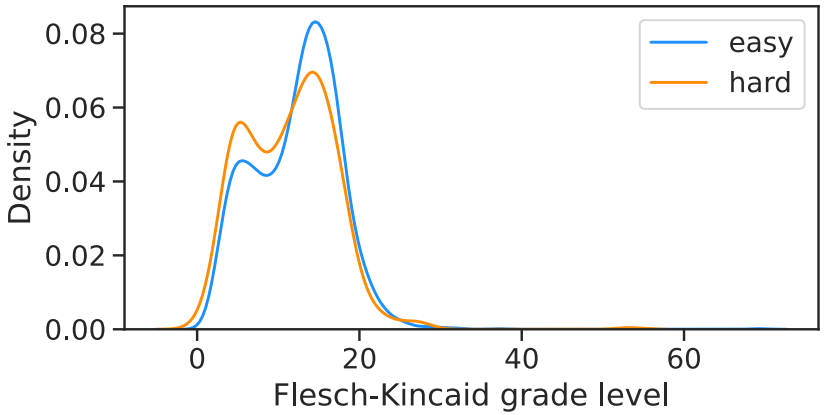
Results: Human–Model Performance Gap

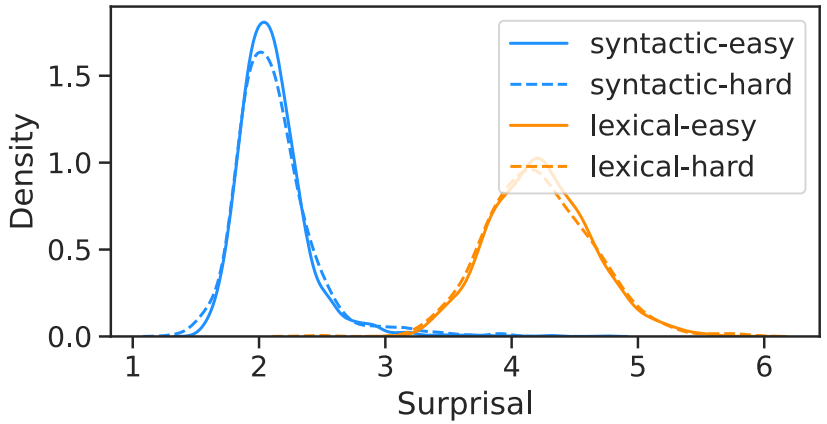


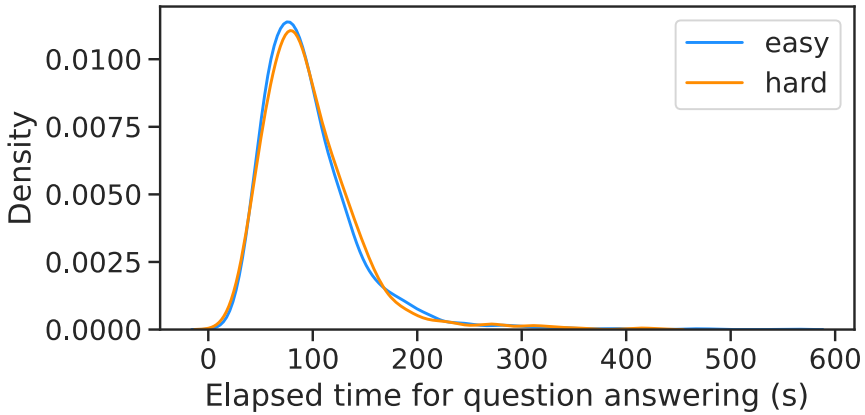
- Small variation across sources (Δ=14.9 ± 3.6)
- Largest gap = MCTest (Children's stories)
  - larger than Gutenberg (adults' stories)!
- Human performance may correspond to the passage difficulty (e.g., MCTest & RACE vs Gutenberg & Slate), but this trend doesn't apply to machine performance

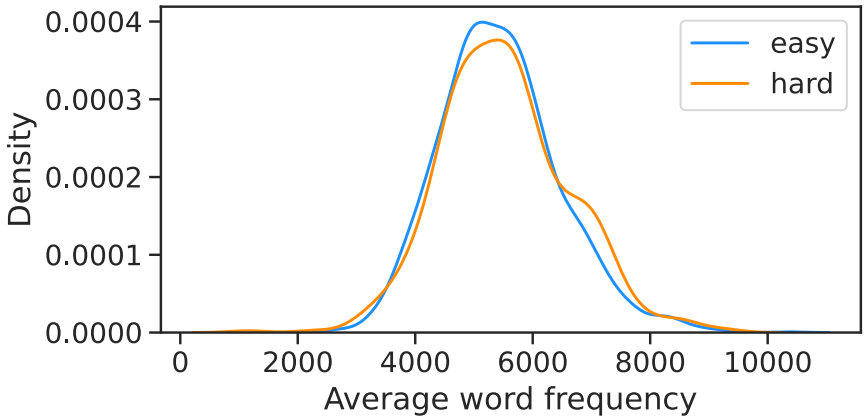
		High-agreement portion				
Source	Method	Human	UniQA	DeBERTa	M-Avg.	Δ
MCTest	Dir.	95.0	71.5	88.2	81.5	**13.5**
	Adv.	**96.5**	27.9	78.6	68.2	**28.3**
	Total	**95.8**	49.3	83.3	74.7	**21.1**
Gutenberg	Dir.	92.8	75.0	88.5	83.4	9.4
	Adv.	87.5	28.3	82.6	72.9	14.6
	Total	90.3	53.1	85.7	78.4	11.9
Slate	Dir.	90.7	74.6	91.7	87.0	3.8
	Adv.	92.9	27.9	76.0	**73.8**	19.1
	Total	91.8	52.6	84.3	**80.8**	11.0
RACE	Dir.	95.4	74.8	90.4	84.6	10.8
	Adv.	94.3	31.0	73.8	67.3	27.0
	Total	94.9	53.3	82.2	76.1	18.8
ReClor	Dir.	**96.9**	79.6	91.1	84.4	12.5
	Adv.	88.8	32.4	74.5	71.3	17.5
	Total	93.2	58.1	83.5	78.5	14.8
Wiki. Sci.	Dir.	95.8	79.0	94.9	87.3	8.5
	Adv.	92.8	29.4	77.2	68.3	24.5
	Total	94.4	56.3	86.8	78.6	15.8
Wiki. Arts	Dir.	91.5	77.0	92.5	**88.1**	3.4
	Adv.	91.4	25.8	75.8	71.7	19.7
	Total	91.5	52.3	84.5	80.2	11.2
All sources	Dir.	94.0	75.9	91.0	85.2	8.8
	Adv.	92.0	29.0	76.9	70.5	21.5
	Total	93.1	53.6	84.3	78.2	14.9

Analysis: Correlation with Linguistic Aspects









Δ = human - system

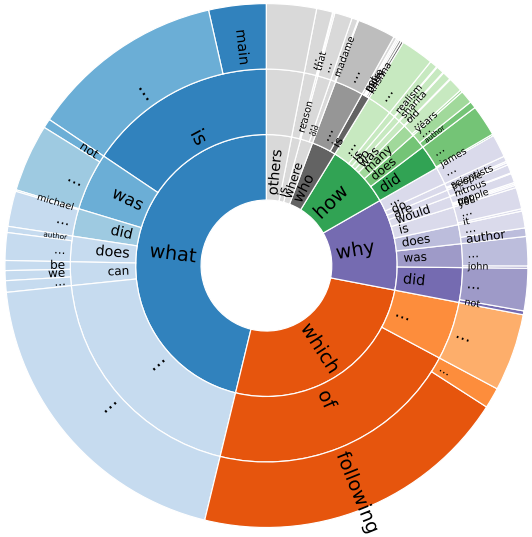
easy: Δ ≤ 20% acc

hard: Δ ≥ 40% acc

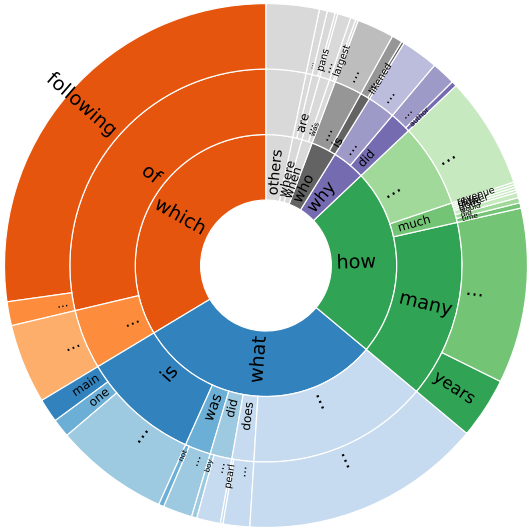
No statistically significant correlations with question difficulty!

Analysis: Question Types

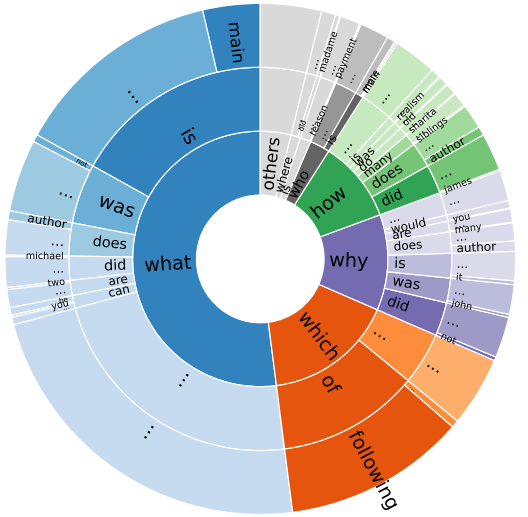
- Hard questions seem to be generic, not specific to given passages (e.g., “which of the following is correct?”)
- Many “how many” questions in Hard
- Questions in Easy are more balanced (because the standard Qs are?)
- This trend is probably because the workers focus on writing specific types (i.e., generic and numeric) of questions in the adversarial data collection



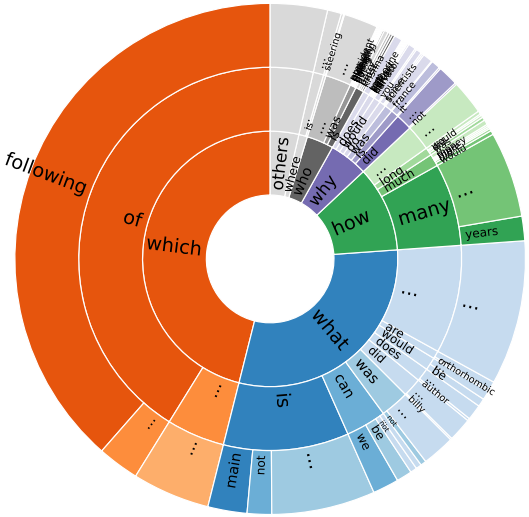
Easy questions



Hard questions



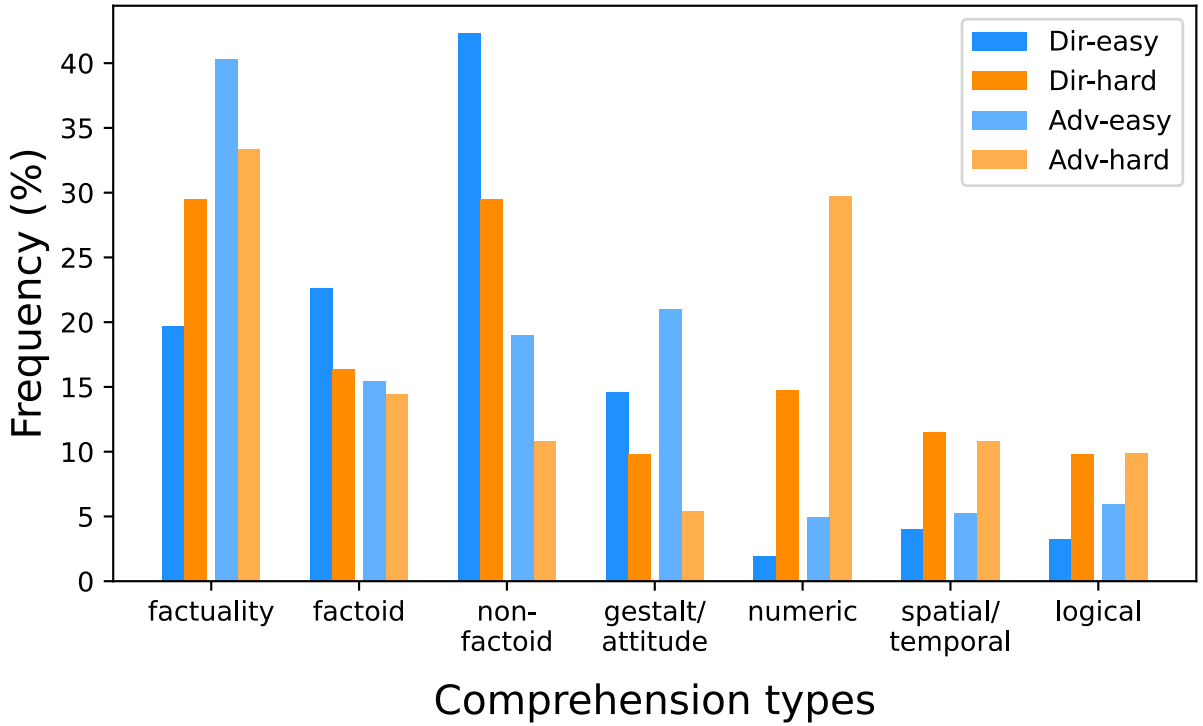
Standard Collection



Adveresarial Collection

Analysis: Comprehension Types vs Difficulty

- We annotated 980 example with seven comprehension types
- Numeric, spatial/temporal, and logical questions appear more often in the hard subset in both collection methods



Analysis: Comprehension Types vs Sources

We observe some trends:

- Technical documents (ReClor & Slate)

  → Logical reasoning questions

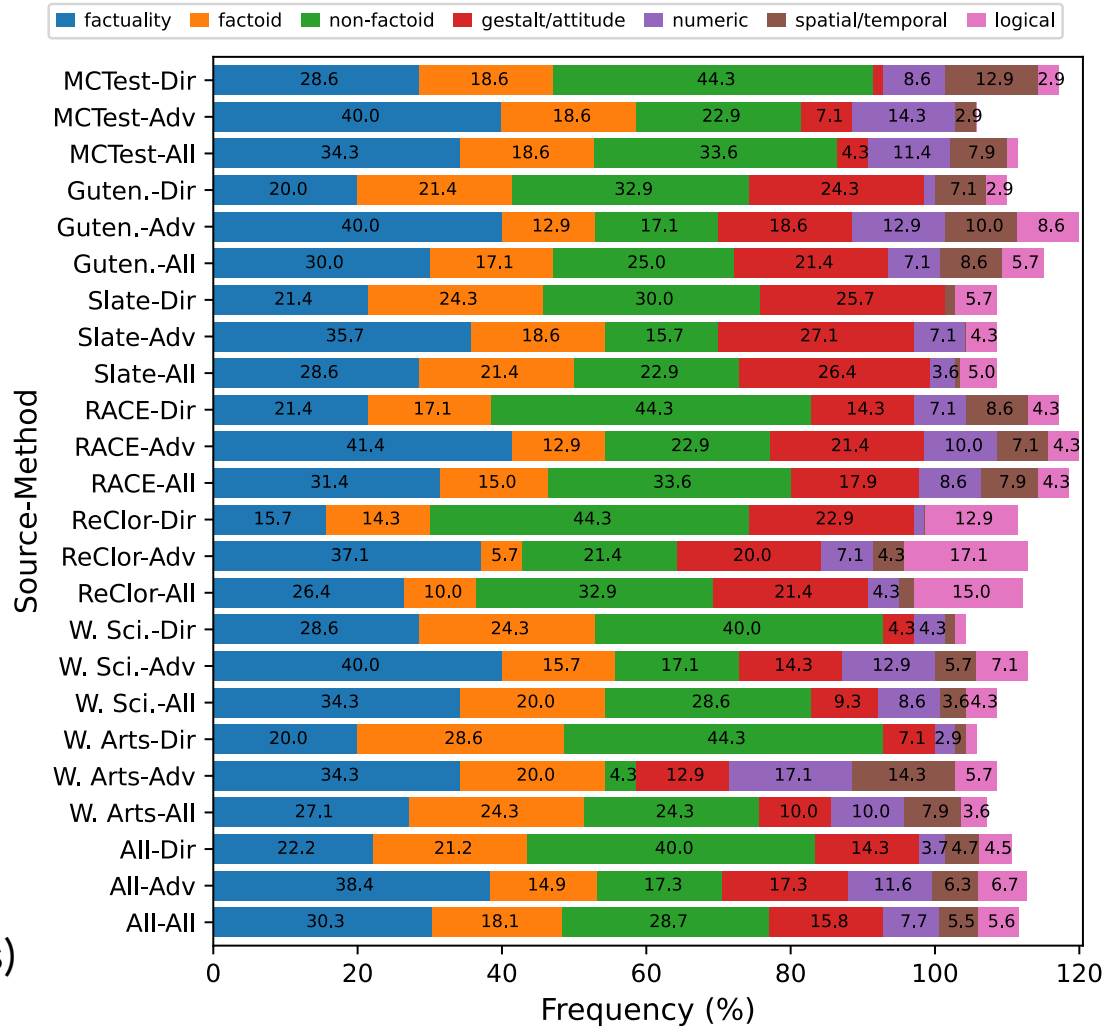
- Subjective or argumentative topics (Gutenberg, Slate, & ReClor)

  → Gestalt/author's attitude questions

- Numbers in passages (MCTest, Wiki arts)

  → Num reasoning in the adv. collection

(Consistent with Kaushik et al. (2021)'s observations)



Summary & Takeaway

- Passage difficulty does not affect question difficulty
- Selecting a diverse set of passages can help ensure a diverse range of reasoning types
- Adversarial data collection has a risk to encourage workers to focus on writing only a few specific types of questions

Our data is available at https://github.com/nii-cl/qa_text_source_comparison