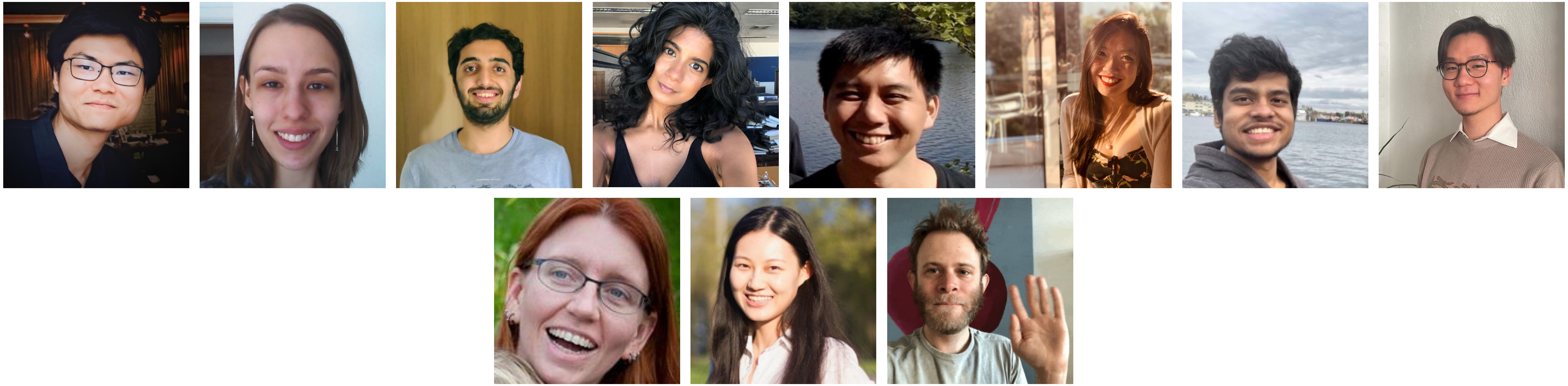
QuALITY: Question Answering with Long Input Texts, Yes!

Richard Yuanzhe Pang*   Alicia Parrish*   Nitish Joshi*   Nikita Nangia   Jason Phang   Angelica Chen
Vishakh Padmakumar   Johnny Ma   Jana Thompson   He He   Samuel R. Bowman



Motivation

- Long document understanding (English)
- Issues with existing work
  - Lack of good-quality long-context datasets: contexts are short; questions are too easy; mostly generation-based datasets (hard to evaluate)
  - Lack of appropriate models (limitation of context size, retrieval, reasoning)

Motivation

- Long document understanding (English)
- Issues with existing work
  - Lack of good-quality long-context datasets: contexts are short; questions are too easy; mostly generation-based datasets (hard to evaluate)
  - Lack of appropriate models (limitation of context size, retrieval, reasoning)
- Goals
  - A high-quality dataset (high human perf; actually measures high-level long-context understanding)
  - Release the dataset including the articles word-by-word instead of URLs (so, need to choose the sources wisely)

Task

Multiple-choice question answering

- Input
  - Article: ~2k—8k tokens; avg 5159 tokens (spaCy tokenization)
    - Source: Gutenberg, Slate magazine articles, and other non-fiction articles
  - Question
  - Four options (1 correct option and 3 distractors)
- Output
  - Option number

story (5159 tokens, avg. 25 min read)

Which is the best representation of Dr. Lessing's worries about his book?

A: He is anxious about the amount of time it will take to revise

✓ B: He is concerned that having to back up his claims could keep him from being objective

C: He is having second thoughts about his qualifications to publish a volume like this

D: He is not sure how he will be able to publish the facts without including the confusing information about the boy

Goal for Each Question

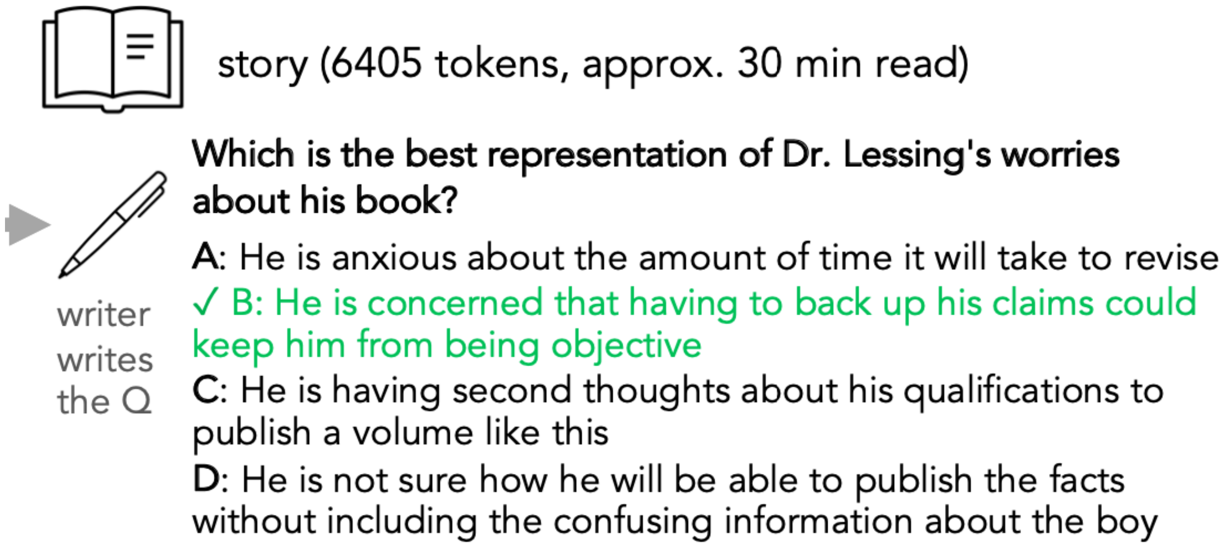
- Correct and unambiguous
- Difficult: rely on more than a few paragraphs of context; good distractors

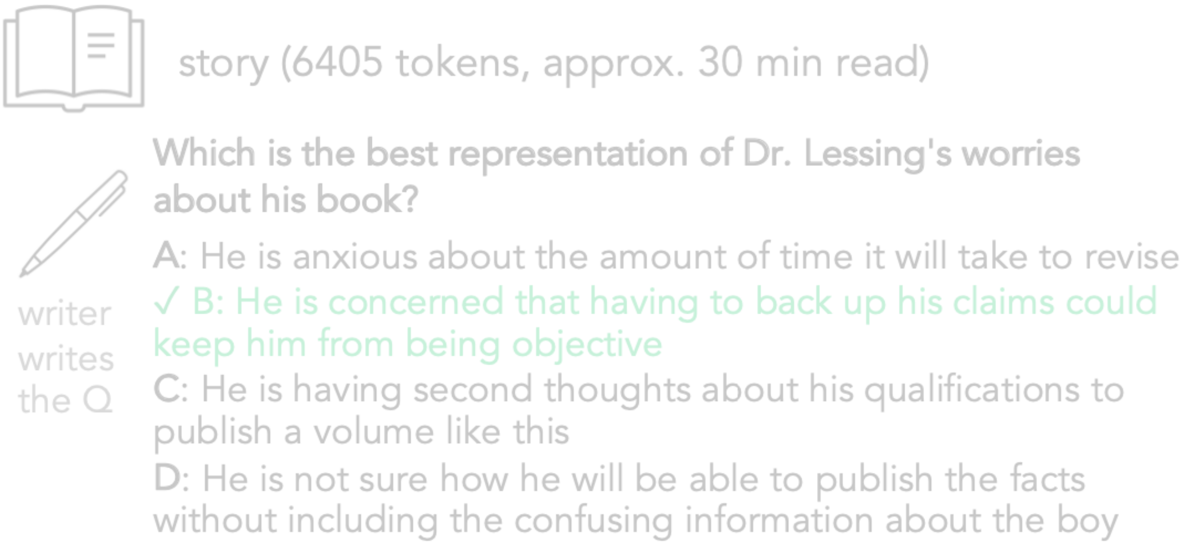
Data Collection Part 1: Writing

Upwork freelancers read the entire article; write 10 multiple-choice questions that are correct+unambiguous and difficult.

Base pay: $12.5 (plus additional fees)

Bonus: $1.2 per question (on avg, writers receive bonus on 42% Qs)





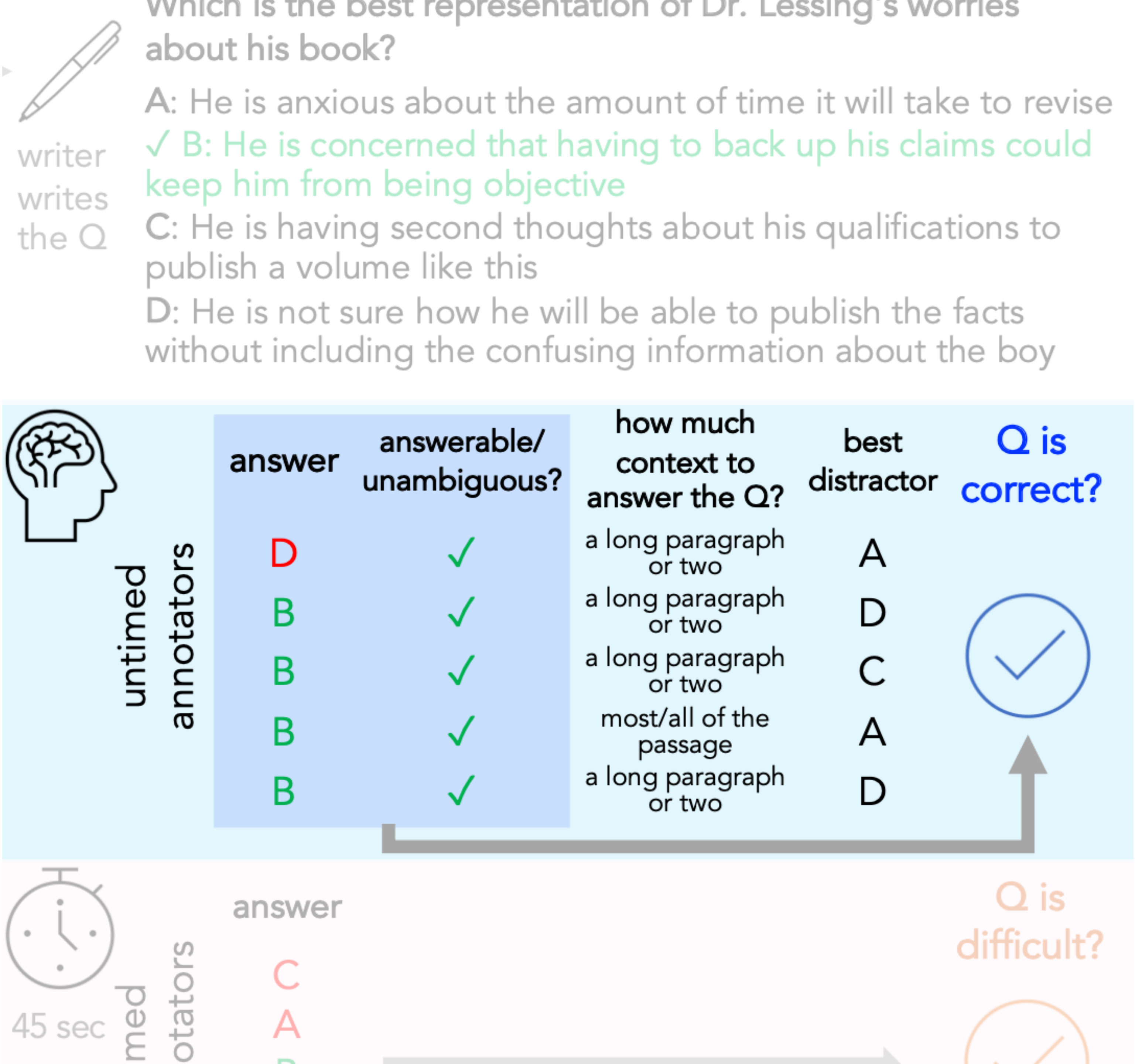
Data Collection Part 2: Speed Validation

(to identify Qs that are too easy)

- Able to quickly identify the answer to a Q; then Q too easy
- Adversarial data collection strategy – but replacing models with humans

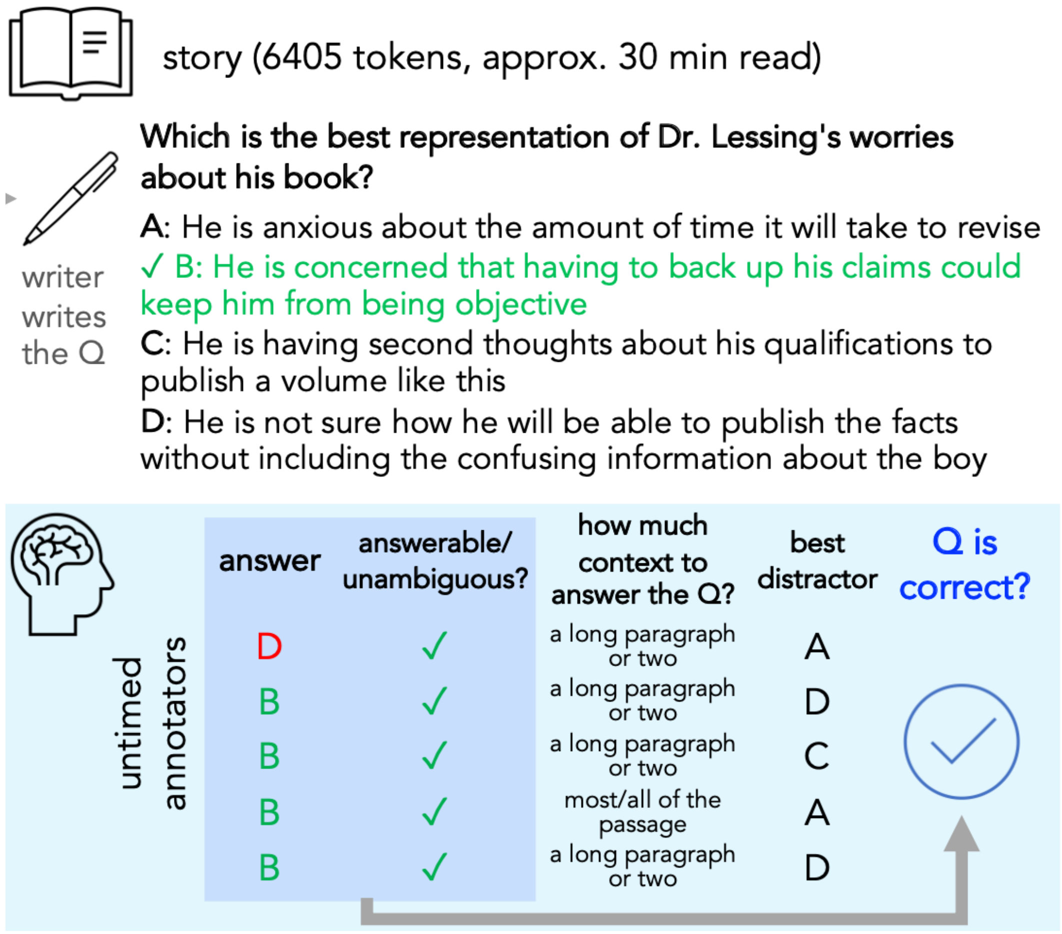
Data Collection Part 3: Untimed Validation

(to ensure correctness and encourage difficulty of Qs)



Dataset

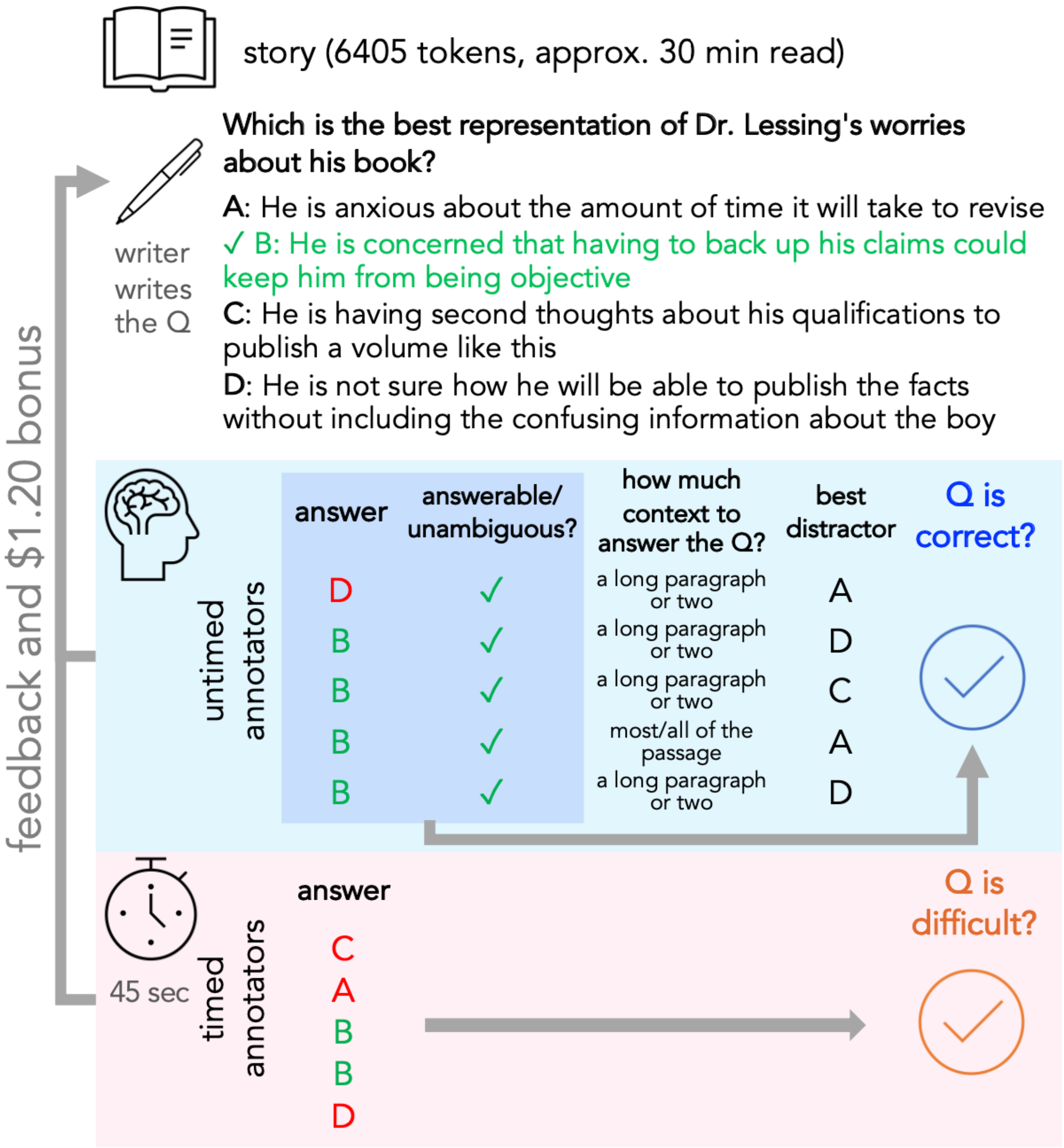
- It is a **correct** question if
  - ≥ 3 out of 5 annotators think that the question is answerable and unambiguous in <u>untimed</u> validation
  - ≥ 3 out of 5 annotators pick the correct option in <u>untimed</u> validation



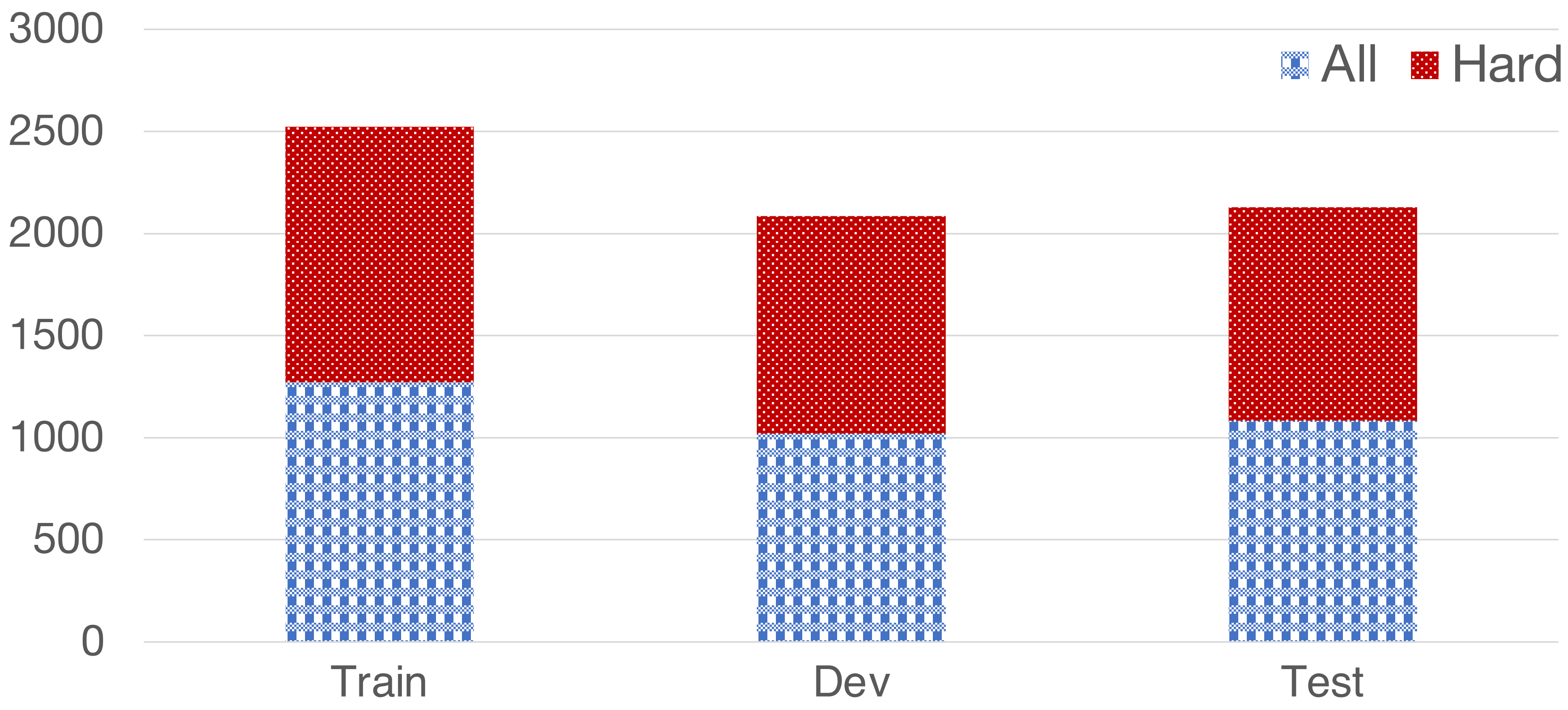
Only correct Qs are included in the dataset

Dataset

- It is a **correct** question if
  - ≥ 3 out of 5 annotators think that the question is answerable and unambiguous in untimed validation
  - ≥ 3 out of 5 annotators pick the correct option in untimed validation
- It is a **correct+hard** question if
  - The above two bullet points
  - ≤ 2 annotators pick the correct option in speed validation



Data Splits



49.9% questions are hard

Reasoning Strategies

What isn't a reason for narrator to be so skeptical of Gorb?
(a) Gorb looked just like an Earthling (b) Gorb was asking for too much money (c) Gorb had no proof to back up his claims (d) he had never heard of Wazzenazz

Why/reason
not/except

Description
Why/reason
Symbolism/interpretation
How/method
Event
Person
Not/except
Relation
Entity
Finish the phrase
Location
Numeric
Object
What if
Duration

Reasoning Strategies

Many questions rely on…

- reasoning about the best description (33.2%)
- determining the correct explanation for why something happened (31.2%)
- the reader making an interpretation or using symbolism (27.8%)

All these types are likely to rely on broader context from the passage.

Description
Why/reason
Symbolism/interpretation
How/method
Event
Person
Not/except
Relation
Entity
Finish the phrase
Location
Numeric
Object
What if
Duration

Modeling

Architecture

- Encoder models: Longformer, RoBERTa, DeBERTaV3
- Encoder-decoder models: Longformer encoder-decoder (LED), T5

Training set (after initializing with pretrained checkpoint of a model above)

- (1) QuALITY; (2) RACE; (3) Intermediate training on RACE→QuALITY

Extraction/retrieval (for non-Longformer models) based on…

- (1) ROUGE-1 recall; (2) fastText; (3) DPR

Architecture

- Encoder models: Longformer, RoBERTa, DeBERTaV3
- Encoder-decoder models: Longformer encoder-decoder (LED), T5

Training set (after initializing with pretrained checkpoint of a model above)

- (1) QuALITY; (2) RACE; (3) Intermediate training on RACE→QuALITY

Extraction/retrieval (for non-Longformer models) based on…

- (1) ROUGE-1 recall; (2) fastText; (3) DPR

Obs 1 (comparing different models):

DeBERTaV3-large generally performs better than other models (given the same training data and the same extraction strategy)

Obs 2 (comparing different training data):

Training on RACE→QuALITY > RACE > QuALITY, in general (given the same extraction strategy and the same model)

	Model	Extraction by R-1	Extraction by fastText	Extraction by DPR
QuALITY	DeBERTaV3-large	46.5 / 39.3	45.5 / 40.2	49.0 / 41.2
RACE	DeBERTaV3-large	52.9 / 43.4	51.2 / 42.4	53.0 / 44.4
RACE -> QuALITY	DeBERTaV3-large	53.8 / 46.3	54.7 / 46.7	55.4 / 46.1

Each cell: full / hard accuracy

Obs 3 (comparing different retrieval strategies):

DPR/fastText performs better than ROUGE-1 recall, in general (given the same model and the same training data)

	Model	**Extraction by R-1**	**Extraction by fastText**	**Extraction by DPR**
QuALITY	DeBERTaV3-large	46.5 / 39.3	45.5 / 40.2	49.0 / 41.2
RACE	DeBERTaV3-large	52.9 / 43.4	51.2 / 42.4	53.0 / 44.4
RACE -> QuALITY	DeBERTaV3-large	53.8 / 46.3	54.7 / 46.7	55.4 / 46.1

Each cell: full / hard accuracy

Obs 4 (full dataset vs. hard subset):

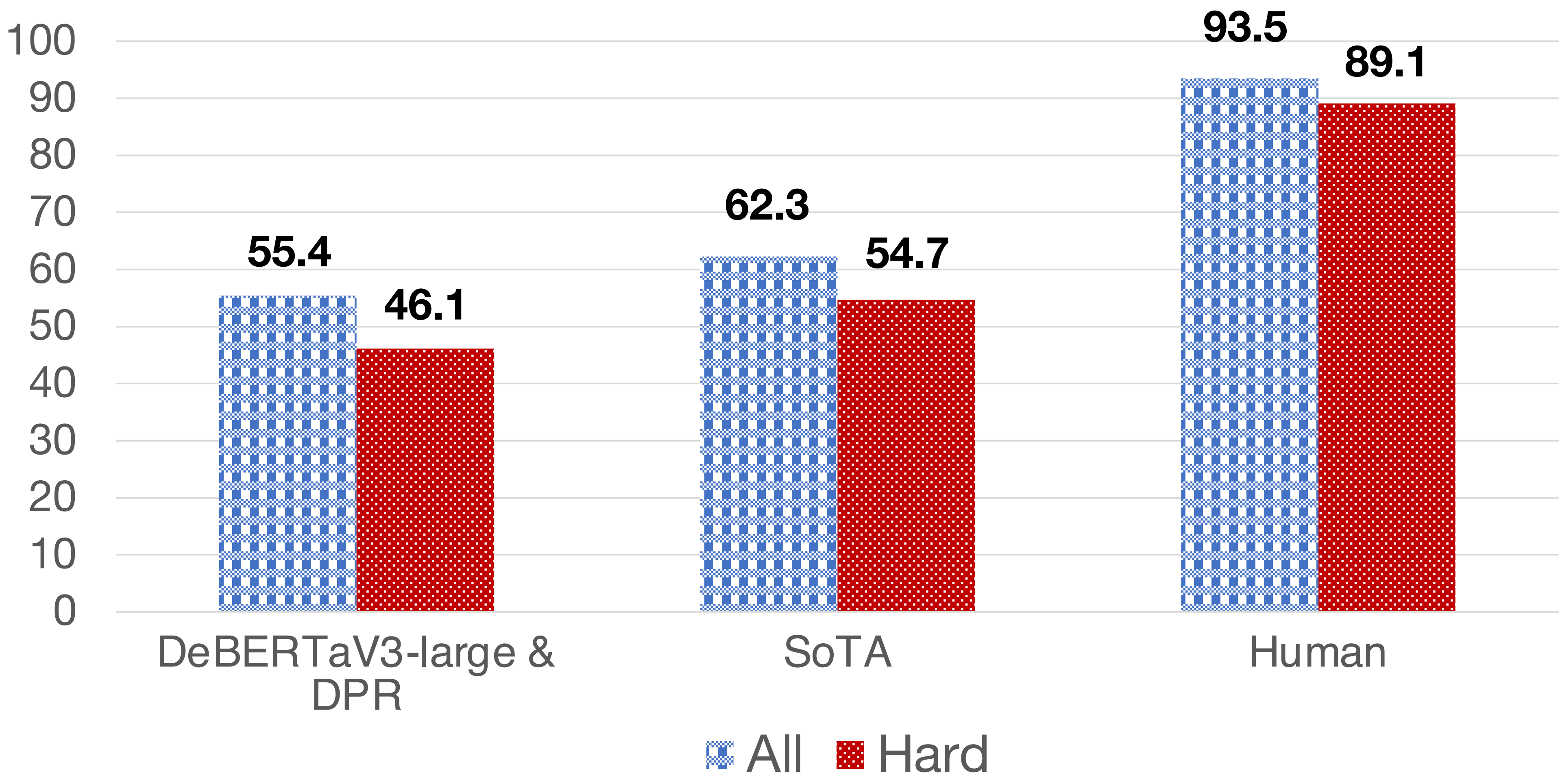
Full dataset performance > hard subset performance

	Model	Extraction by R-1	Extraction by fastText	Extraction by DPR
QuALITY	DeBERTaV3-large	46.5 / 39.3	45.5 / 40.2	49.0 / 41.2
RACE	DeBERTaV3-large	52.9 / 43.4	51.2 / 42.4	53.0 / 44.4
RACE -> QuALITY	DeBERTaV3-large	53.8 / 46.3	54.7 / 46.7	55.4 / 46.1

Each cell: full / hard accuracy

Obs 5 (machine vs. human):

Machine performance < human performance



Leaderboard!

Ours: https://nyu-mll.github.io/quality/

SAT-style score:

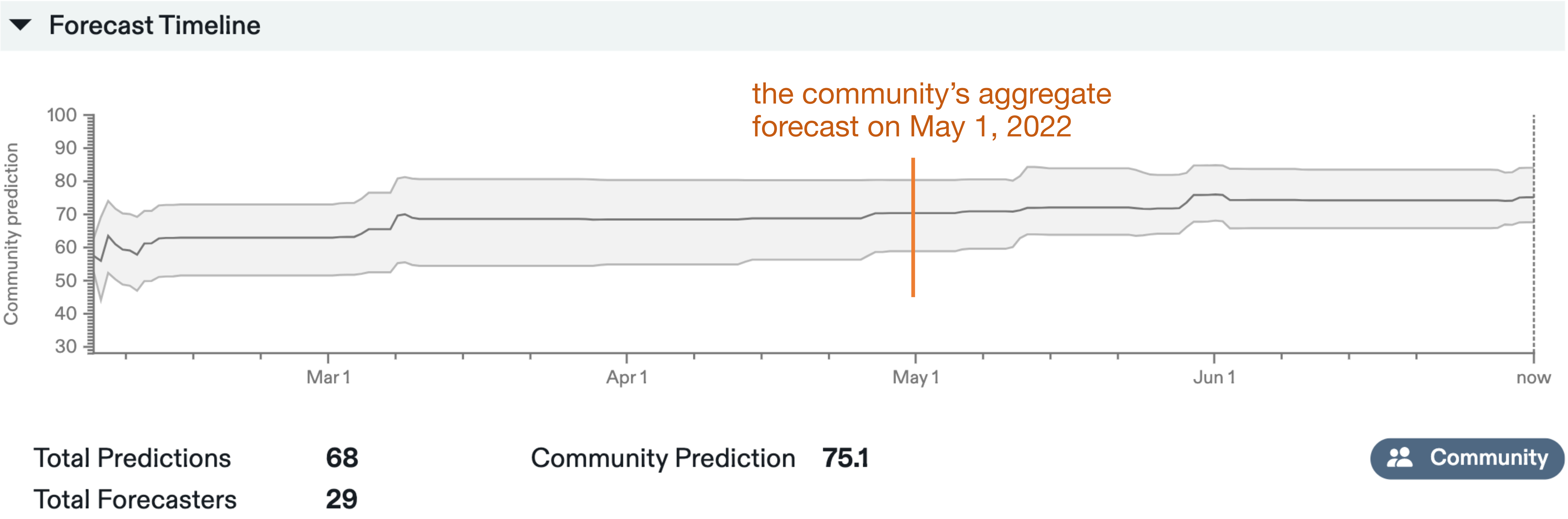
(# correct answers – 1/3 * # incorrect answers + 0 * # abstained answers) / # questions * 100

	Model name		Paper	Code	Accuracy		SAT-style score	
					Test set	Hard subset	Test set	Hard subset
0 2021/12	Human annotators *New York University*	description	🗎		93.5	89.1	91.4	85.4
1 2022/05	CoLISA: DPR & DeBERTaV3-large architecture plus contrastive learning & in-sample attention *Anonymous (temporary)*	description			62.3	54.7	49.7	39.6
2 2022/04	CoLISA: DPR & DeBERTaV3-large architecture & contrastive learning *Anonymous (temporary)*	description			62.1	54.3	49.5	39.1
3 2021/12	Baseline model: DPR retrieval using questions & DeBERTaV3-large with intermediate training on RACE *New York University*	description	🗎	</>	55.4	46.1	40.5	28.1

Forecast! (by elifland on Metaculus)

https://www.metaculus.com/questions/9628/question-answering-on-long-texts-by-2025/

What will be the best non-human SAT-style score on the hard subset of the QuALITY dataset by January 1, 2025? (screenshot taken on July 4, 2022)



Forecast! (by elifland on Metaculus)

https://www.metaculus.com/questions/9628/question-answering-on-long-texts-by-2025/
What will be the best non-human SAT-style score on the hard subset of the QuALITY dataset by January 1, 2025?

https://www.metaculus.com/questions/9629/question-answering-on-long-texts-by-2030/
What will be the best non-human SAT-style score on the hard subset of the QuALITY dataset by January 1, 2030?

https://www.metaculus.com/questions/9630/question-answering-on-long-texts-by-2040/
What will be the best non-human SAT-style score on the hard subset of the QuALITY dataset by January 1, 2040?

Leaderboard!

SCROLLS (Shaham et al., 2022): `www.scrolls-benchmark.com`

Date	Model	Contributors	#Params	Input Length	Score (Average)	GovRep (R1/R2/RL)	SumScr (R1/R2/RL)	QMSum (R1/R2/RL)	Qspr (F1)	Nrtv (F1)	QALT (EM-T/H)	CNLI (EM)
04/27/2022	LongT5 XL	LongT5	3B	16K	41.89	54.7/28.2/30.2	35.8/9.6/21.1	34.9/11.8/23.5	53.1	29.3	46.0/42.1	88.2
04/28/2022	LongT5 Large	LongT5	770M	16K	40.47	54.2/27.8/29.8	35.6/9.2/21.2	35.1/12.0/23.3	52.3	27.2	40.6/38.6	87.3
04/28/2022	LongT5 Base	LongT5	220M	16K	38.22	53.5/27.3/29.3	34.8/9.6/21.1	33.9/11.0/22.8	46.6	23.0	37.9/36.6	85.6
03/14/2022	UL2	Google Research	20B	2K	37.87	53.6/26.1/28.8	32.9/7.8/19.4	31.1/8.5/20.4	37.6	24.2	45.8/40.7	88.7
01/01/2022	LED Base	SCROLLS team	162M	16K	29.16	56.2/26.6/28.8	24.2/4.5/15.4	25.1/6.7/18.8	26.6	18.5	25.8/25.4	71.5
01/01/2022	BART Base	SCROLLS team	139M	1K	29.01	47.9/18.6/22.7	27.2/4.9/16.7	30.2/8.7/20.7	26.3	15.4	26.0/25.9	77.4
01/07/2022	Naive	SCROLLS team	-	-	19.35	45.3/17.9/20.8	19.6/1.8/11.0	14.2/2.0/9.3	3.4	1.5	25.2/26.1	66.0

Recap

Crowdsourcing: writing (Upwork), speed/timed validation (MTurk), untimed validation (MTurk)

Modeling

- Architecture: DeBERTaV3, LED, T5, long T5, etc.
- Training set: QuALITY; RACE; RACE -> QuALITY
- Extraction/retrieval: ROUGE-1 recall; fastText; DPR

Future modeling

- Better architectures, transfer learning strategies, extraction/retrieval strategies?
- Explicitly modeling entity relationships for better story understanding?