Trick Me If You Can: Human-in-the-loop Generation of Adversarial Examples for Question Answering

Eric Wallace, Pedro Rodriguez, Shi Feng,
Ikuya Yamada, and Jordan Boyd-Graber
University of Maryland, Studio Ousia

Experts + Computers collaborate to create adversarial examples

Why? Adversarial examples highlight model vulnerabilities

How? Our UI provides model interpretations and predictions

Predictions →

Machine Guesses [Update All]

#	Guess	Confidence
1	Sorting algorithm	0.54
2	Permutation	0.05
3	Fisher–Yates shuffle	0.05
4	Quicksort	0.05
5	Radix sort	0.03

Settings
Automatic Updates Every **5** Words
[Modify Existing Question] [New Question]

Sorting algorithm [Submit]

Random permutations of the objects involved in this process are generated until the correct permutation is found in the bogo type of this process. If the values of the elements involved come from a known finite set like the integers, then the radix one of these processes is appropriate. Other methods include a divide-and-conquer algorithm using a pivot value, and, in another type of this process, "rabbits" are put in the correct place very quickly, while "turtles" find their way through the list slowly. For 10 points, quick and bubble are types of which algorithms that arrange the elements of a list in ascending or descending order.

QANTA Buzz on: Other methods include a divide-and-conquer

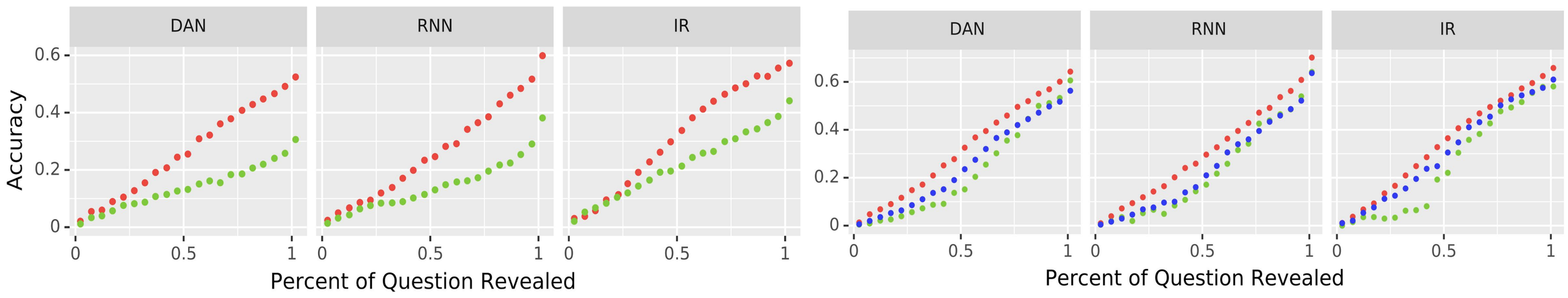
Evidence for [Sorting algorithm] [More Evidence]

Your Question	Evidence
Random permutations of the objects involved in this process are generated until the correct permutation is found in the bogo type of this process.	is called the radix type, and another type of this process randomly places elements and checks if... (Quiz Bowl)
Other methods include a divide-and-conquer **Buzz** algorithm using a pivot value, and, in another type of this process, "rabbits" are put in the correct place very quickly, while "turtles" find their way through the list slowly.	- and - conquer algorithm and the selection of a pivot element and which runs in big O of n log n time... (Quiz Bowl)

← Interpretations

Quizbowl Trivia experts create adversarial examples with the UI

Models struggle on questions



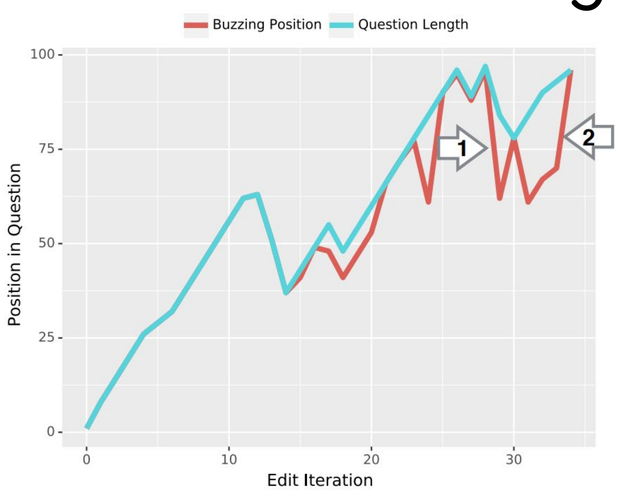
Adversarial question types

Composing Seen Clues	15%
Logic & Calculations	5%
Multi-Step Reasoning	25%
Novel Clues	26%
Entity Type Distractors	7%
Paraphrases	38%
Total Questions	1213

This number is one hundred fifty more than the number of Spartans at Thermopylae.

taking of one's own life — self-inflicted method of death

The interface guides writers



~~The BioLIp database stores data on the interaction of these species with proteins.~~
Examples of these molecules with C2 symmetry can increase enantioselectivity, as in their Josiphos variety. . .
Prediction: Ion (✗) → Ligand (✓) [1]

Examples of these ~~molecules~~ species with C2 symmetry can increase enantioselectivity, as in their Josiphos variety. . .
Prediction: Ligand (✓) → Ion (✗) [2]

trickme.qanta.org