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

Anonymous TACL submission

Abstract

Adversarial evaluation is a promising paradigm to stress test a model's ability to understand natural language. While past approaches expose superficial patterns learned by models, the resulting adversarial examples are limited in complexity and diversity. We propose a human-in-the-loop adversarial generation process, where human authors are guided to break models. We aid the authors by providing model interpretations in an interactive user interface which helps to expose a system's inner workings. We apply this generation framework to a question answering task called Quizbowl, leveraging trivia enthusiasts to craft adversarial questions. We validate the resulting questions via live human–computer tournaments, showing that although they appear ordinary for human players, the questions systematically stump both neural and information retrieval models. The adversarial questions cover diverse phenomena, spanning multi-hop reasoning to entity type distractors, exposing remaining challenges in robust question answering.

1 Introduction

Proponents of machine learning claim human parity on tasks like reading comprehension (Yu et al., 2018) and commonsense inference (Devlin et al., 2018). Despite these successes, many evaluations neglect that computers solve NLP tasks in a fundamentally different way than humans.

Models can succeed without developing "true" language understanding, instead learning superficial patterns from crawled (Chen et al., 2016) or manually annotated datasets (Kaushik and Lipton, 2018; Gururangan et al., 2018). Thus, recent work stress tests models via adversarial evaluation: elucidating a system's capabilities by exploiting its weaknesses (Jia and Liang, 2017; Belinkov and Glass, 2019). Unfortunately, while adversarial evaluation reveals simplistic model failures (Ribeiro et al., 2018; Mudrakarta et al., 2018), exploring more complex failure patterns requires human involvement (Figure 1): automatically modifying natural language while maintaining example validity is difficult. Hence, the diversity of adversarial examples is often severely restricted.

Instead, we leverage trivia enthusiasts—who write questions for academic competitions—to create diverse adversarial examples that stump existing Quizbowl question answering models. This human–computer hybrid approach uses human creativity to generate more diverse adversarial examples than previous work. To aid writers in crafting adversarial examples, we create a user interface that presents model interpretations and predictions (Section 3).

The adversarially-authored test set is nonetheless easier than regular questions for humans (Section 4), but strong QA models only achieve 60% of their original performance (Section 5). We also host live human vs. computer matches, where models typically defeat top human teams, and observe spectacular model failures on adversarially-authored questions.

Analyzing the adversarial edits uncovers phenomena that humans can solve but computers cannot (Section 6), validating that our framework allows creative, targeted adversarial edits (Section 7). Our resulting adversarial dataset presents a fun, challenging, and diverse resource for future QA research: a system that masters it will demonstrate more robust language understanding.

2 Adversarial Evaluation for NLP

Adversarial examples (Szegedy et al., 2013) often reveal model failures better than traditional test sets. However, automatic adversarial generation is

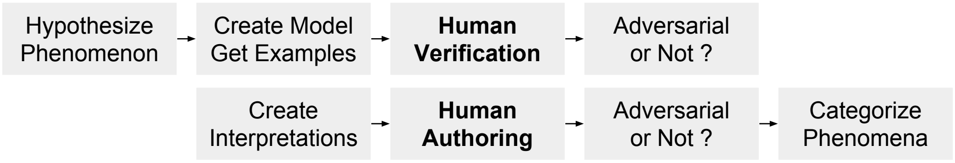


Figure 1: Adversarial evaluation in NLP typically focuses on a specific phenomenon (e.g., word replacements) and then generates the corresponding examples (top). Consequently, adversarial examples are limited to the diversity of what the underlying generative model can produce, and also require downstream *human evaluation* to ensure validity. Our setup (bottom) instead has *human-crafted* examples, using human–computer collaboration to craft adversarial examples with greater diversity.

tricky for NLP (e.g., by replacing words) without changing an example's meaning or invalidating it.

Recent work side-steps this by focusing on simple transformations which preserve meaning. For instance, Ribeiro et al. (2018) generate adversarial perturbations such as replacing *What has* → *What's*. Other minor perturbations such as typos (Belinkov and Bisk, 2018), adding distractor sentences (Jia and Liang, 2017; Mudrakarta et al., 2018), or character replacements (Ebrahimi et al., 2018) preserve meaning while degrading model performance.

Generative models can discover more adversarial perturbations, but require verifying examples through post-hoc human evaluation. For example, neural paraphrase or language models can generate syntax modifications (Iyyer et al., 2018), plausible captions (Zellers et al., 2018), or NLI premises (Zhao et al., 2018). These methods improve example-level diversity but still target a specific phenomenon, e.g., rewriting question syntax.

Furthermore, existing adversarial perturbations are restricted to sentences—not the paragraph inputs of Quizbowl and other tasks—due to challenges in long-text generation. For instance, syntax paraphrase networks (Iyyer et al., 2018) applied to Quizbowl only yield valid paraphrases 3% of the time (Appendix A).

2.1 Putting a Human in the Loop

Instead, we task human authors with *adversarial writing* of questions: generating examples which break a specific QA system but are still answerable by humans. We expose model predictions and interpretations to question authors, who see what changes would confuse the model.

The user interface makes the adversarial writing process interactive and model-driven, in contrast to adversarial examples written independently of a model (Ettinger et al., 2017). The result is an adversarially-authored dataset that explicitly exposes a model's limitations by design.

Human-in-the-loop generation can replace or aid model-based adversarial generation approaches. Creating interfaces and interpretations is often easier than designing and training generative models for specific domains. In domains where adversarial generation is feasible, human creativity can reveal which tactics automatic approaches can later emulate. Model-based and human-in-the-loop generation approaches can also be combined by training models to mimic human adversarial edit history, using the relative merits of both approaches.

3 Our QA Testbed: Quizbowl

The "gold standard" of academic competitions between universities and high schools is Quizbowl. Unlike QA formats such as Jeopardy! (Ferrucci et al., 2010), Quizbowl questions are designed to be interrupted: questions are read to two competing teams and whoever knows the answer first interrupts the question and "buzzes in".

The answers to Quizbowl questions are typically well-known entities. In the QA community (Hirschman and Gaizauskas, 2001), this is called "factoid" question answering: the entities come from a relatively closed set of possible answers (e.g., Wikipedia page titles).

This style of play requires questions to be structured "pyramidally" (Jose, 2017): questions start with difficult clues and get progressively easier. These questions are carefully crafted by the authors to allow the most knowledgeable player to answer first. A question on Paris that begins "this capital of France" would test reaction speed, not knowledge; thus, skilled question authors arrange the clues so players will recognize them with increasing probability (Figure 2).

> The protagonist of this opera describes the future day when her lover will arrive on a boat in the aria "Un Bel Di" or "One Beautiful Day". The only baritone role in this opera is the consul Sharpless who reads letters for the protagonist, who has a maid named Suzuki. That protagonist blindfolds her child Sorrow before stabbing herself when her lover B. F. Pinkerton returns with a wife. For 10 points, name this Gia4o Puccini opera about an American lieutenant's affair with the Japanese woman Cio-Cio San.
> **Answer**: Madama Butterfly

Figure 2: An example Quizbowl question. The question becomes progressively easier (for humans) to answer later on; thus, more knowledgeable players can answer after hearing fewer clues. Our adversarial writing process ensures that the clues, especially early ones, also challenge computers.

3.1 Known Exploits of Quizbowl Questions

Like most QA datasets, Quizbowl questions are written for *humans*. Unfortunately, the heuristics that question authors use to select clues do not always apply to computers. For example, humans are unlikely to memorize every song in every opera by a particular composer. This, however, is trivial for a computer. In particular, a simple QA system easily solves the example in Figure 2 from seeing the reference to "Un Bel Di". Other questions contain uniquely identifying "trigger words" (Harris, 2006). For example, "martensite" only appears in questions on steel. For these examples, a QA system needs to understand no additional information other than an if–then rule.

One might wonder if this means that factoid QA is thus an uninteresting, nearly solved research problem. However, some Quizbowl questions are fiendishly difficult for computers. Many questions have intricate coreference patterns (Guha et al., 2015), require reasoning across multiple types of knowledge, or involve complex wordplay. If we can isolate and generate questions with these difficult phenemona, "simplistic" factoid QA can become extremely non-trivial.

3.2 Models and Datasets

We conduct two rounds of adversarial writing. In the first, authors attack a traditional Information Retrieval (IR) system. The model is an IR system distributed as the baseline for a NIPS 2017 shared task on Quizbowl (Boyd-Graber et al., 2018) based on ElasticSearch (Gormley and Tong, 2015). The inverted index is built using approximately 60,000 questions from the shared task.

In the second round, authors attack either the IR model or a neural QA model. The neural model is a bidirectional RNN using the gated recurrent unit architecture (Cho et al., 2014). The model treats Quizbowl as classification and predicts the answer entity from a sequence of words represented as 300-dimensional GloVe embeddings (Pennington et al., 2014). Both models in this round are trained using an expanded dataset of approximately 110,000 Quizbowl questions. We expand the dataset to incorporate a more diverse answer set (25,000 entities versus 11,000 in round one).

3.3 Interpreting Quizbowl Models

To help write adversarial questions, we expose what the model is thinking to authors. We interpret models using saliency heat maps: each word of the question is highlighted based on its importance to the model's prediction (Ribeiro et al., 2016).

For the neural model, word importance is the decrease in prediction probability when a word is removed (Li et al., 2016; Wallace et al., 2018). We focus on gradient-based approximations (Simonyan et al., 2014; Montavon et al., 2017), which are computationally efficient.

To interpret a model prediction on an input sequence of n words $\mathbf{w} = \langle \boldsymbol{w}_1, \boldsymbol{w}_2, \ldots \boldsymbol{w}_n \rangle$, we approximate the classifier f with a linear function of w_i derived from the first-order Taylor expansion. The importance of w_i, with embedding $\boldsymbol{v}_i$, is the derivative of f with respect to the one-hot vector:

$$\frac{\partial f}{\partial w_i} = \frac{\partial f}{\partial \boldsymbol{v}_i}\frac{\partial \boldsymbol{v}_i}{\partial w_i} = \frac{\partial f}{\partial \boldsymbol{v}_i} \cdot \boldsymbol{v}_i. \tag{1}$$

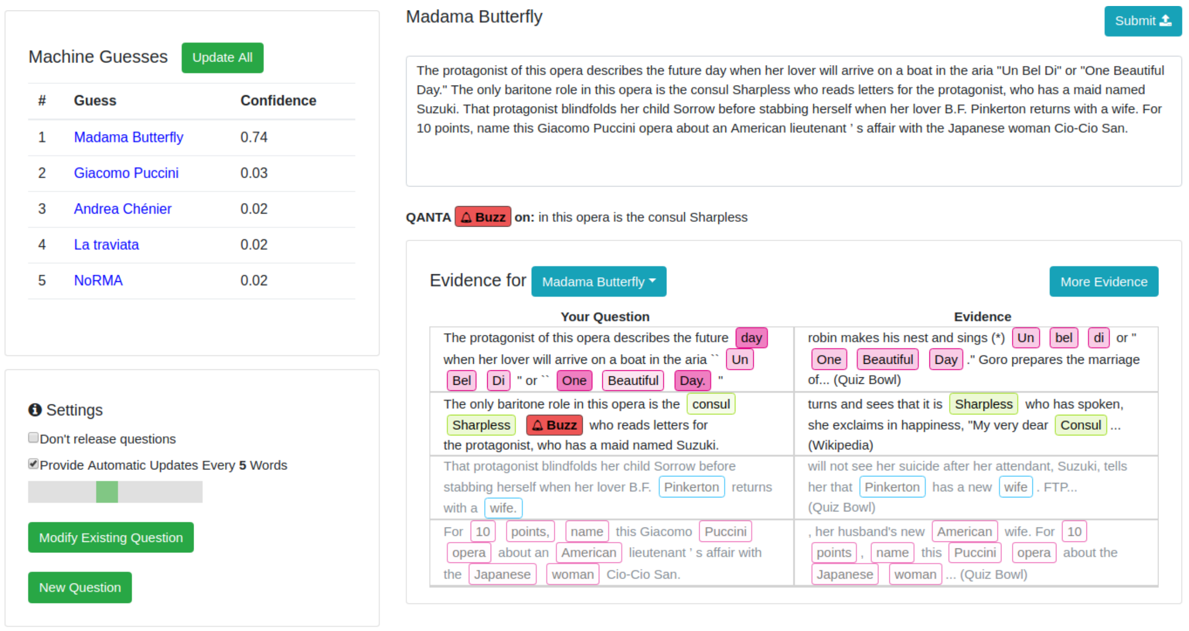
This simulates the change in prediction probability when a word's embedding is set to the zero vector—i.e., approximating word removal—and is a common interpretation method for NLP (Ebrahimi et al., 2018; Wallace et al., 2018).

For the IR model, we use the ElasticSearch Highlight API (Gormley and Tong, 2015), which provides word importance scores based on query matches from the inverted index.

3.4 Adversarial Writing Interface

The author interacts with either the IR or RNN model through a user interface[1] (Figure 3). The

[1] Code for the interface, models, and interpretations available after blind review.



#	Guess	Confidence
1	Madama Butterfly	0.74
2	Giacomo Puccini	0.03
3	Andrea Chénier	0.02
4	La traviata	0.02
5	NoRMA	0.02

□ Don't release questions

☑ Provide Automatic Updates Every **5** Words

The protagonist of this opera describes the future day when her lover will arrive on a boat in the aria "Un Bel Di" or "One Beautiful Day." The only baritone role in this opera is the consul Sharpless who reads letters for the protagonist, who has a maid named Suzuki. That protagonist blindfolds her child Sorrow before stabbing herself when her lover B.F. Pinkerton returns with a wife. For 10 points, name this Giacomo Puccini opera about an American lieutenant ' s affair with the Japanese woman Cio-Cio San.

QANTA Buzz **on:** in this opera is the consul Sharpless

Your Question	Evidence
The protagonist of this opera describes the future day when her lover will arrive on a boat in the aria `` Un Bel Di " or `` One Beautiful Day. "	robin makes his nest and sings (*) Un bel di or " One Beautiful Day ." Goro prepares the marriage of... (Quiz Bowl)
The only baritone role in this opera is the consul Sharpless Buzz who reads letters for the protagonist, who has a maid named Suzuki.	turns and sees that it is Sharpless who has spoken, she exclaims in happiness, "My very dear Consul ... (Wikipedia)
That protagonist blindfolds her child Sorrow before stabbing herself when her lover B.F. Pinkerton returns with a wife.	will not see her suicide after her attendant, Suzuki, tells her that Pinkerton has a new wife . FTP... (Quiz Bowl)
For 10 points, name this Giacomo Puccini opera about an American lieutenant ' s affair with the Japanese woman Cio-Cio San.	, her husband's new American wife. For 10 points , name this Puccini opera about the Japanese woman ... (Quiz Bowl)

Figure 3: The author inputs a question (top right), the QA system provides guesses (left), and explains why it's making those guesses (bottom right). The author can then adapt their question to "trick" the model.

author writes their question in the upper right while the model's top five predictions (*Machine Guesses*) appear in the upper left. If the top prediction is the right answer, the interface indicates where in the question the model is first correct. The author's goal is to cause the model to be incorrect or to delay the correct answer position as much as possible.[2] The words of the current question are highlighted using the applicable interpretation method in the lower right (*Evidence*). We do not enforce time restrictions or require questions to be adversarial: if the author fails to break the system, they are free to "give up" and submit any question.

The interface continually updates as the author writes. We track the question edit history to identify recurring model failures (Section 6) and understand how interpretations guide authors' (Section 7).

3.5 Question Authors

We focus on members of the Quizbowl community: they have deep trivia knowledge and craft questions for Quizbowl tournaments (Jennings, 2006). We award prizes for questions read at live human–computer matches (Section 5.3).

The question authors are familiar with the standard format of Quizbowl questions (Lujan and Teitler, 2003). The questions follow a common paragraph structure, are well edited for grammar, and finish with a simple "give-away" clue. These constraints benefit the adversarial writing process as it is very clear what constitutes a difficult but valid question. Thus, our examples go beyond surface level "breaks" such as character noise (Belinkov and Bisk, 2018) or syntax changes (Iyyer et al., 2018). Rather, questions are difficult because of their semantic content (examples in Section 6).

3.6 How an Author Writes a Question

To see how an author might write a question with the interface, we walk through an example of writing a question's first sentence. The author first selects the answer to their question from the training set—Johannes Brahms—and begins:

> Karl Ferdinand Pohl showed this **composer** some pieces on which this composer's Variations on a Theme by Haydn were based.

The QA system *buzzes* (i.e., it has enough information to interrupt and answer correctly) after "composer". The author sees that the name "Karl Ferdinand Pohl" appears in Brahms' Wikipedia page and avoids that specific phrase, describing Pohl's position instead of naming him directly:

[2] The authors want normal Quizbowl questions which humans can easily answer by the very end. For popular answers, (e.g., Australia or Suez Canal), writing novel final give-away clues is difficult. We thus expect models to often answer correctly by the very end of the question.

Science	17%
History	22%
Literature	18%
Fine Arts	15%
Religion, Mythology, Philosophy, and Social Science	13%
Current Events, Geography, and General Knowledge	15%
Total Questions	1213

Table 1: The topical diversity of the questions in the adversarially-authored dataset based on a random sample of 100 questions.

> This composer was given a theme called "Chorale St. Antoni" by the archivist of the Vienna Musikverein, which could have been written by Ignaz Pleyel.

The QA system now incorrectly thinks the answer is Frédéric Chopin. The author continues this process to create entire questions the model cannot solve.

4 A New Adversarially-Authored Dataset

Our adversarial dataset consists of 1213 questions with 6,541 sentences across diverse topics (Table 1).[3] 807 questions were written against the IR system and 406 against the neural model by 115 unique authors. We plan to hold twice-yearly competitions to continue data collection.

4.1 Validating Questions with Quizbowlers

We validate that the questions are not of poor quality or too difficult for humans. We first automatically filter out invalid questions based on length, the presence of vulgar statements, or repeated submissions (including re-submissions from the Quizbowl training or evaluation data).

We next host a human-only Quizbowl event using intermediate and expert players (former and current collegiate players). We select sixty adversarially-authored questions and sixty unreleased high school national championship questions both with the same number of questions per category (list of categories in Table 1).

To answer a Quizbowl question, a player interrupts the question: the earlier the better. To capture this dynamic, we record both the average answer position (as a percentage of the question seen, lower is better) and answer accuracy. We randomly shuffle the baseline and adversarially-authored questions, read them to players, and record these two metrics.

The adversarially-authored questions are on average *easier* for humans than the regular test questions. For the adversarially-authored set, humans buzz with 41.6% of the question remaining and an accuracy of 89.7%. On the baseline questions, humans buzz with 28.3% of the question remaining and an accuracy of 84.2%. The difference in accuracy between the two types of questions is not significantly different ($p = 0.16$ using Fisher's exact test), but the buzzing position is earlier for adversarially-authored questions ($p = 0.0047$ for two-sided t-test). We expect human performance to be comparable on the questions not played, as all questions went through the same submission process and post-processing. We further explore the human-perceived difficulty of the adversarially-authored questions in Section 5.3.

5 Computer Experiments

This section evaluates QA systems on the adversarially-authored questions. We test three models: the IR and RNN models shown in the interface, as well as a Deep Averaging Network (Iyyer et al., 2015, DAN) to test the transferability of the adversarial questions. We break our study into two rounds. The first round consists of adversarially-authored questions written against the IR system (Section 5.1); the second round questions target both the IR and RNN systems (Section 5.2).

Finally, we also hold live competitions that pit the state-of-the-art Studio Ousia model (Yamada et al., 2018) against human teams (Section 5.3).

5.1 Round One: IR Adversarial Questions Transfer To All Models

The first round of adversarially-authored questions target the IR model, and are significantly harder for the IR, RNN, and DAN models (Figure 4). For example, the DAN's accuracy drops from 54.1% to 32.4% on the full question (60% of original performance).

For both adversarially-authored and original test questions, the early clues are difficult to answer (accuracy about 10% through 25% of the question). However, during the middle third of the questions, where buzzes in Quizbowl most frequently occur, the accuracy on original test questions rises significantly quicker than the adversarially-authored ones. For both questions, the accuracy rises to-

[3] Dataset available after blind review.

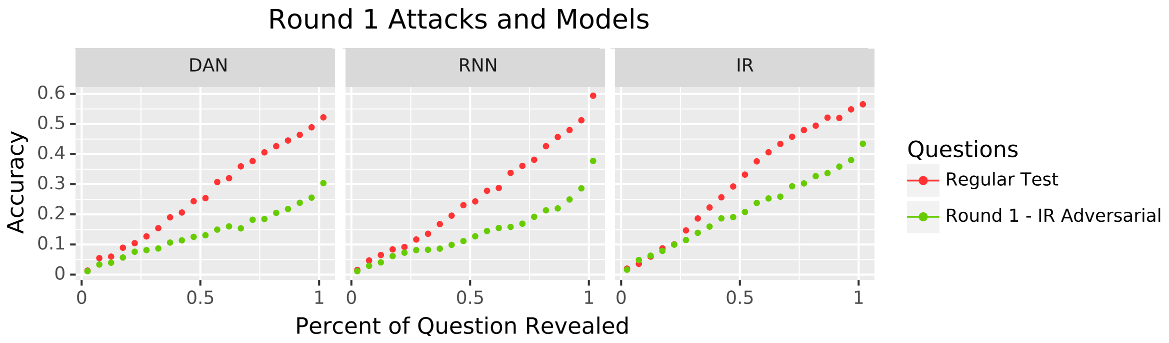


Figure 4: The first round of adversarial writing attacks the IR model. Like regular test questions, adversarially-authored questions begin with difficult clues that trick the model. However, the adversarial questions are significantly harder during the crucial middle third of the question.

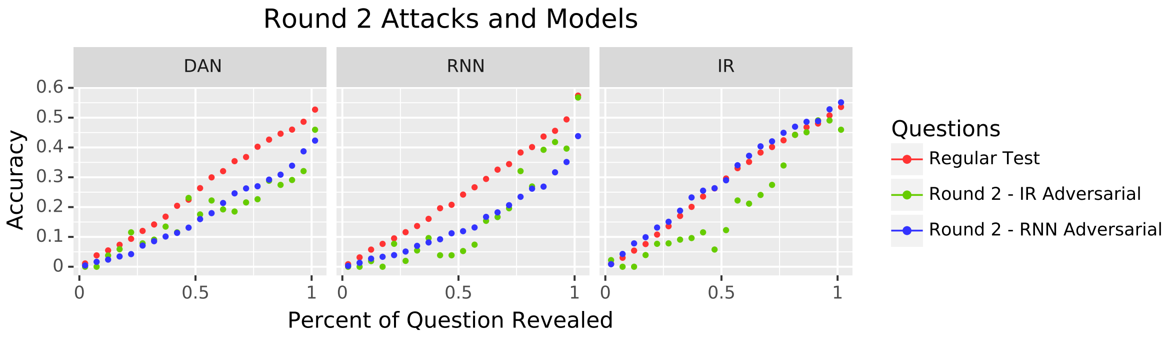


Figure 5: The second round of adversarial writing attacks both IR and RNN models. The questions targeted against the IR system (*IR Adversarial*) degrade the performance of all models. However, the reverse does not hold: the IR model is robust to the questions written to fool the RNN (*RNN Adversarial*).

wards the end as the clues get simpler and become "give-aways".

5.2 Round Two: RNN Adversarial Questions are Brittle

The adversarial questions written in the first round are hard for all tested models. In the second round, authors additionally attack an RNN model. The models tested in the second round are trained on a larger dataset (Section 3.2).

A similar trend holds for IR adversarial question in round two (Figure 5): a question that tricks the IR system also fools the two neural models (i.e., the adversarial examples transfer). For example, the DAN model was never targeted adversarially but had substantial accuracy decreases in both rounds.

However, this does not hold for questions written adversarially against the RNN model. On these questions, the neural models struggle but the IR model is largely unaffected (Figure 5, right).

5.3 Humans vs. Computer, Live!

In the offline setting (i.e., no pressure to "buzz" before an opponent) models demonstrably struggle on the adversarial questions. But, what happens in standard Quizbowl: live, head-to-head games?

We run two live humans vs. computer matches. The first match uses IR adversarial questions in a forty question, tossup-only Quizbowl format. We pit a human team of national-level Quizbowl players against the Studio Ousia model (Yamada et al., 2018), the current state-of-the-art Quizbowl system. The model combines neural, IR, and knowledge graph components (details in Appendix B), and won the 2017 NIPS shared task, defeating a team of expert humans 475–200 on regular Quizbowl test questions. The developers of this system had no access to the adversarially-authored questions. Although the team at our live event was comparable to the NIPS 2017 team, the tables were turned: the human team won handedly 300–30.

Our second live event is significantly larger in scale: seven human teams play against models on over 400 questions written adversarially against the RNN model. The human teams range in ability from high school Quizbowl players to national-level teams. The models are based on either IR or neural methods. Despite a few close games

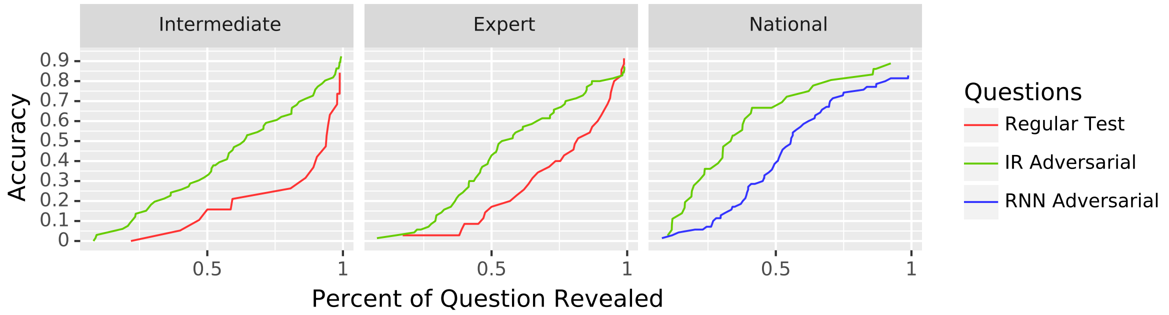


Figure 6: Humans find adversarially-authored question about as difficult as normal questions regardless of whether they are rusty weekend warriors (*Intermediate*), active players (*Expert*), or some of the best trivia players in the world (*National*).

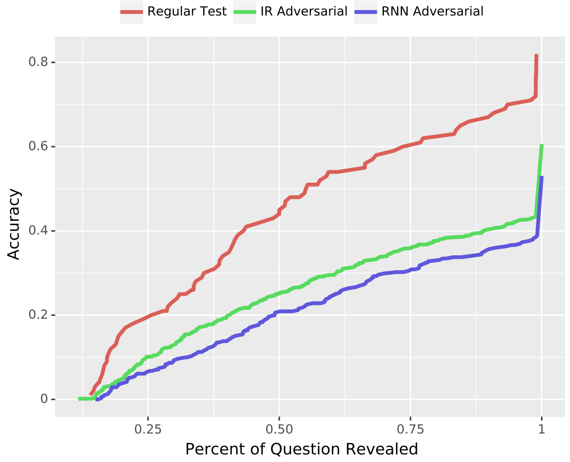


Figure 7: The accuracy of the state-of-the-art Ousia model degrades on the adversarially-authored questions despite never being targeted by the adversarial authors. This verifies that our findings generalize beyond the RNN and IR models.

between the weaker human teams and the models, humanity prevailed in every match.[4]

Figures 6–7 summarize the live match results for humans and the Ousia model, respectively. The accuracy trend between humans and models is considerably different. Human accuracy on both regular and adversarial questions rises very quickly in the *last* 50% of the question (curves in Figure 6). In essence, the "give-away" clues at the end of questions are easy for humans to answer.

On the other hand, models on *regular* test questions do well in the *first* 50%, i.e., the "difficult" clues for humans are easier for models (*Regular Test* curve in Figure 7). However, models, like humans, struggle to answer the adversarial questions in the first 50% of the question.

6 What Makes Adversarially-authored Questions Hard?

This section analyzes the adversarially-authored questions to locate the source of their difficulty.

6.1 Quantitative Differences in Questions

One possible source of difficulty is scarcity: the answers to adversarial questions rarely appear in the training set. However, this is not the case; the mean number of training examples per answer (e.g., George Washington) is 14.9 for the adversarial questions versus 16.9 for the regular test data.

Another explanation for question difficulty is limited "overlap" with the training data, i.e., models cannot match n-grams from the training clues. We measure the proportion of test n-grams that also appear in training questions with the same answer (Table 2). The overlap is roughly equal for unigrams but surprisingly higher for bigrams in adversarial questions. The adversarial questions are also shorter and have fewer NEs on average. However, the proportion of question words which are NEs is roughly equivalent.

One noticeable difference between the questions written against the IR system and the ones written against the RNN model is the drop in NEs. The decrease is much higher for IR adversarial questions, which may explain their generalization: the RNN is more sensitive to changes in phrasing, while the IR system is more sensitive to specific words.

6.2 Categorizing Adversarial Phenomena

We next qualitatively analyze the adversarially-authored questions. We manually inspect the author edit logs, classifying the questions into six different phenomena that fall into two broad categories. We report the relative frequency of each phenomenon

[4] Videos of matches available after blind review.

	Adversarial	**Regular**
Unigram overlap	0.40	0.37
Bigram overlap	0.08	0.05
Longest n-gram overlap	6.73	6.87
Average NE overlap	0.38	0.46
IR Adversarial	0.35	
RNN Adversarial	0.44	
Total Words	107.1	133.5
Total NE	9.1	12.5

Table 2: The adversarially-authored questions have similar n-gram overlap to the regular test questions. However, there is a decrease in the overlap of the named entities (NE) with those from the training data for IR Adversarial questions.

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

Table 3: A breakdown of the challenging phenomena in the adversarially-authored dataset.

from a random sample of 100 questions, double counting questions into multiple phenomena when applicable (Table 3).

6.2.1 Category 1: Reasoning

The first question category requires reasoning about known clues (Table 4).

Composing Seen Clues: In these questions, several entities that have a first-order relationship to the correct answer are given. The system must then triangulate the correct answer by "filling in the blank". For example, in the first question of Table 4, the place of death of the entity is given. The training data contains a clue about the place of death (The Battle of the Thames) reading "though stiff fighting came from their Native American allies under Tecumseh, who died at this battle". The system must connect these two clues to answer.

Logic & Calculations: These adversarial questions require applying mathematical or logical operators. For example, the training data contains a clue about the Battle of Thermopylae reading "King Leonidas and 300 Spartans died at the hands of the Persians". The second question in Table 4 requires adding 150 to the number of Spartans.

Multi-Step Reasoning: This question type requires multiple reasoning steps between entities. For example, in the last question of Table 4, a model needs to make a step from the "I Have A Dream" speech to the Lincoln Memorial and another step to reach Abraham Lincoln.

6.2.2 Category 2: Distracting Clues

The second category consists of circumlocutory clues (Table 5).

Paraphrases: A common adversarial modification is to paraphrase clues to remove exact n-gram matches from the training data. This renders an IR system useless but also hurts neural models. Many of the adversarial paraphrases go beyond syntax-only changes (e.g., the first row of Table 5).

Entity Type Distractors: Whether explicit or implicit in a model, one key component for QA is determining the answer type of the question. Authors take advantage of this by providing clues that cause the model to select the wrong answer type. For example, in the second question of Table 5, the "lead-in" clue implies the answer may be an actor. The RNN model answers Don Cheadle in response despite previously seeing the Bill Clinton "playing a saxophone" clue.

Novel Clues: Some adversarially-authored questions are hard not because of phrasing or logic but because our models simply have not seen these clues. These questions are relatively easy for users to create: users can add *Novel Clues* that—because they are not uniquely associated with any answer—tend to confuse the models. While not as linguistically interesting, this does add clues that are not captured by Wikipedia or previous questions, improving the diversity of the dataset. For example, adding clues about literary criticism (Hardwick, 1967; Watson, 1996) to a question about Lillian Hellman's The Little Foxes: "Ritchie Watson commended this play's historical accuracy for getting the price for a dozen eggs right—ten cents—to defend against Elizabeth Hardwick's contention that it was a sentimental history." This creates more interesting, varied questions for humans and an incentive for models to use diverse sources of information beyond Wikipedia.

Question	Prediction	Answer	Phenomenon
This man, who died at the Battle of the Thames, experienced a setback when his brother Tenskwatawa's influence over their tribe began to fade.	Battle of Tippecanoe	Tecumseh	Composing Seen Clues
This number is one hundred fifty more than the number of Spartans at Thermopylae.	Battle of Thermopylae	450	Logic & Calculations
A building dedicated to this man was the site of the "I Have A Dream" speech.	Martin Luther King Jr.	Abraham Lincoln	Multi-Step Reasoning

Table 4: Snippets from adversarially-authored questions show examples of reasoning about existing evidence. *Answer* displays the correct answer (all models were incorrect). For these examples, connecting the training and adversarially-authored clues is simple for humans but difficult for models.

Set	Question	Prediction	Phenomenon
Training	Name this sociological phenomenon, the *taking of one's own life*.	Suicide	Paraphrase
Adversarial	Name this *self-inflicted method of death*.	Arthur Miller	
Training	Clinton played the *saxophone on The Arsenio Hall Show*.	Bill Clinton	Entity Type Distractor
Adversarial	He was edited to appear in the film "Contact"... For ten points, name this American president who played the *saxophone on an appearance on the Arsenio Hall Show*.	Don Cheadle	

Table 5: Snippets from adversarially-authored questions show the difficulty in retrieving previously seen evidence. *Training* questions indicate relevant snippets from the training data. *Prediction* displays the RNN model's answer prediction (always correct on Training, always incorrect on Adversarial).

Novel clues have different effects on IR and neural models: while IR models largely ignore them, novel clues can lead neural models astray. For example, on a question about Tiananmen Square, the RNN model buzzes on the clue "World Economic Herald". However, adding a novel clue about "the history of shaving" renders the brittle RNN unable to buzz on the "World Economic Herald" clue that it was able to recognize before.[5] This helps to explain why adversarially-authored questions written against the RNN do not stump IR models.

7 How Do Interpretations Help?

This section explores how interpretations help to guide adversarial authors. We analyze the question edit log, which reflects how an author modifies a question given the current model interpretation.

A direct edit of the highlighted words often creates an adversarial example (e.g., Figure 8). Figure 9 shows a more intricate example. The left plot shows the *Question Length*, as well as the position where the model is first correct (*Buzzing Position*, lower is better performance). We show two adversarial edits. In the first (**1**), the author removes the beginning of the question, which makes the question *easier* for the model (buzz position decreases). The author counteracts this in the second edit (**2**), where they use the interpretation to craft a small, targeted modification that breaks the IR model.

However, models are not always this brittle. In Figure 10 (Appendix C), the interpretation fails to aid an adversarial attack against the RNN model. At each step, the author uses the highlighted words as a guide to edit targeted portions of the question yet fails to trick the model. The author gives up and submits their relatively non-adversarial question.

7.1 Interviews With Adversarial Authors

We also interviewed the adversarial authors who attended our live events. Multiple authors agreed that identifying oft-repeated "stock" clues was the interface's most useful feature. As one author explained, "There were clues that I wrote which I did not think were stock clues but were later revealed

[5] The "history of shaving" is a tongue-in-cheek name for a poster displaying the hirsute intellectual giants of Communist thought. It goes from the bearded Marx and Engels, to the mustachioed Lenin and Stalin, and finally the clean-shaven Mao.

One of these concepts . . . a **Hyperbola** is a type of, for ten points, what shapes made by passing a **plane** through a namesake solid, ~~that also includes the **ellipse**, **parabola**~~? whose area is given by one-third Pi r squared times height?
Prediction: Conic Section (✓) → Sphere (✗)

Figure 8: The interpretation successfully aids an attack against the IR system. The author removes the phrase containing the words "ellipse" and "parabola", which are highlighted in the interface (shown in bold). In its place, they add a phrase which the model associates with the answer sphere.

to be." In particular, the author's question about the Congress of Vienna used a lead-in clue about "Kraków becoming a free city," which the model immediately recognized.

Another interviewee was Jordan Brownstein,[6] a national Quizbowl champion and one of the best active players, who felt that computer opponents were better at questions that contained direct references to battles or poetry. He also explained how the different writing styles used by each Quizbowl author increases the difficulty of questions for both humans and models. The interface's evidence panel allows authors to read existing clues which encourages unique stylistic choices.

8 Related Work

New datasets often allow for a finer-grained analysis of a linguistic phenomenon, task, or genre. The LAMBADA dataset (Paperno et al., 2016) tests a model's ability to understand the broad contexts present in book passages, while the Natural Questions corpus (Kwiatkowski et al., 2019) combs Wikipedia for answers to questions that users trust search engines to answer (Oeldorf-Hirsch et al., 2014). Other work focuses on natural language inference, where challenge examples highlight existing model failures (Wang et al., 2018; Glockner et al., 2018; Naik et al., 2018). Our work is unique in that we use human adversaries to expose model weaknesses, which provides a diverse set of phenomena (from paraphrases to multi-hop reasoning) that models cannot solve.

Other work explores specific limitations of NLP systems. Rimell et al. (2009) show that parsers struggle on test examples with unbounded dependencies. A closely related work to ours is Ettinger et al. (2017) who also use human adversaries. Unlike their Build-it Break-it setting, we use interpretation methods to facilitate attacks in an interactive, collaborative manner. Moreover, we have a ready-made audience of "breakers" who are motivated and capable of generating adversarial examples.

The collaborative nature of the adversarial writing process relates to broader investigations into the complementary abilities of humans and computers. For instance, "centaur" chess teams comprised of both a human and computer are often stronger than any human or computer alone (Case, 2018). In Starcraft, humans devise high-level "macro" strategies, while computers are superior at executing fast and precise "micro" actions (Vinyals et al., 2017). In NLP, computers aid simultaneous human interpreters (He et al., 2016) at remembering forgotten information or translating unfamiliar words.

Finally, recent approaches to adversarial evaluation of NLP models (Section 2) typically target one phenomenon (e.g., syntactic modifications) and complement our human-in-the-loop approach.

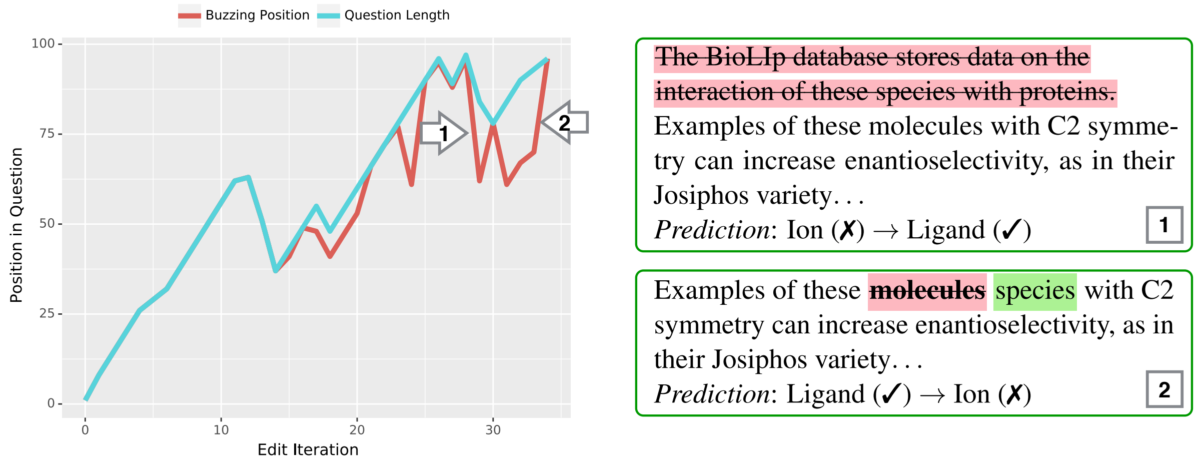
9 Conclusion

One of the challenges of machine learning is knowing why systems fail. This work brings together two threads that attempt to answer this question: visualizations and adversarial examples. Visualizations underscore the capabilities of existing models, while adversarial examples—crafted with the ingenuity of human experts—show that these models are still far from matching human prowess. Our adversarial writing framework increases not only the difficulty but also the diversity of QA data through human–computer collaboration.

Our experiments with both neural and IR methodologies show that QA models still struggle with synthesizing clues, handling distracting information, and adapting to unfamiliar data. Our adversarially-authored dataset is only the first of many iterations (Ruef et al., 2016): as models improve, future adversarially-authored datasets can elucidate the limitations of next-generation QA systems.

While we focus on QA, our procedure is applicable to other NLP settings where there is (1) a pool of talented authors who (2) write text with specific goals. Quizbowl is advantageous as it naturally combines both. Future research can look to craft adversarially-authored datasets for other tasks.

[6] https://www.qbwiki.com/wiki/Jordan_Brownstein



~~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

Figure 9: The *Question Length* and the position where the model is first correct (*Buzzing Position*, lower is better performance) are shown as a question is written. In (**1**), the author makes a mistake by removing a sentence that makes the question easier for the IR model. In (**2**), the author uses the interpretation, replacing the interface's highlighted word (shown in bold) "molecules" with "species" to trick the model.

References

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *Proceedings of the International Conference on Learning Representations*.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. In *Transactions of the Association for Computational Linguistics*.

Jordan Boyd-Graber, Shi Feng, and Pedro Rodriguez. 2018. *Human-Computer Question Answering: The Case for Quizbowl*. Springer.

Nicky Case. 2018. How To Become A Centaur. *Journal of Design and Science*.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the cnn/daily mail reading comprehension task. *Proceedings of the Association for Computational Linguistics*.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of Empirical Methods in Natural Language Processing*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the Association for Computational Linguistics*.

Allyson Ettinger, Sudha Rao, Hal Daumé III, and Emily M. Bender. 2017. Towards linguistically generalizable NLP systems: A workshop and shared task. In *In Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*.

David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, Nico Schlaefer, and Chris Welty. 2010. Building Watson: An Overview of the DeepQA Project. *AI Magazine*, 31(3).

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the Association for Computational Linguistics*.

Clinton Gormley and Zachary Tong. 2015. *Elasticsearch: The Definitive Guide*. " O'Reilly Media, Inc.".

Anupam Guha, Mohit Iyyer, Danny Bouman, and Jordan Boyd-Graber. 2015. Removing the training wheels: A coreference dataset that entertains humans and challenges computers. In *North American Association for Computational Linguistics*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Elizabeth Hardwick. 1967. The little foxes revived. *The New York Review of Books*, 9(11).

Bob Harris. 2006. *Prisoner of Trebekistan: A Decade in Jeopardy!*

He He, Jordan Boyd-Graber, and Hal Daumé III. 2016. Interpretese vs. translationese: The uniqueness of human strategies in simultaneous interpretation. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Lynette Hirschman and Rob Gaizauskas. 2001. Natural language question answering: The view from here. *Natural Language Engineering*, 7(4):275–300.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the Association for Computational Linguistics*.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Ken Jennings. 2006. *Brainiac: adventures in the curious, competitive, compulsive world of trivia buffs*. Villard.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of Empirical Methods in Natural Language Processing*.

Ike Jose. 2017. The craft of writing pyramidal quiz questions: Why writing quiz bowl questions is an intellectual task.

Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of Empirical Methods in Natural Language Processing*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Rhinehart, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, et al. 2019. Natural questions: a benchmark for question answering research.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

Paul Lujan and Seth Teitler. 2003. Writing good quizbowl questions: A quick primer.

Grégoire Montavon, Wojciech Samek, and Klaus-Robert MÃijller. 2017. Methods for interpreting and understanding deep neural networks. *arXiv preprint*, abs/1706.07979.

Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. 2018. Did the model understand the question? In *Proceedings of the Association for Computational Linguistics*.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of International Conference on Computational Linguistics*.

Anne Oeldorf-Hirsch, Brent Hecht, Meredith Ringel Morris, Jaime Teevan, and Darren Gergle. 2014. To search or to ask: the routing of information needs between traditional search engines and social networks. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 16–27. ACM.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the Association for Computational Linguistics*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of Empirical Methods in Natural Language Processing*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging nlp models. In *Proceedings of the Association for Computational Linguistics*.

Laura Rimell, Stephen Clark, and Mark Steedman. 2009. Unbounded dependency recovery for parser evaluation. In *Proceedings of Empirical Methods in Natural Language Processing*.

Andrew Ruef, Michael Hicks, James Parker, Dave Levin, Michelle L. Mazurek, and Piotr Mardziel. 2016. Build it, break it, fix it: Contesting secure development. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Proceedings of the International Conference on Learning Representations*.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. In *Proceedings of the International Conference on Learning Representations*.

Oriol Vinyals, Timo Ewalds, Sergey Bartunov, Petko Georgiev, Alexander Sasha Vezhnevets, Michelle Yeo, Alireza Makhzani, Heinrich Küttler, John Agapiou, Julian Schrittwieser, John Quan, Stephen Gaffney, Stig Petersen, Karen Simonyan, Tom Schaul, Hado van Hasselt, David Silver, Timothy P. Lillicrap, Kevin Calderone, Paul Keet, Anthony Brunasso, David Lawrence, Anders Ekermo, Jacob Repp, and Rodney Tsing. 2017. Starcraft II: A new challenge for reinforcement learning. *arXiv preprint arXiv:1708.04782*.

Eric Wallace, Shi Feng, and Jordan Boyd-Graber. 2018. Interpreting neural networks with nearest neighbors. In *EMNLP 2018 Workshop on Analyzing and Interpreting Neural Networks for NLP*.

Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.05922*.

Ritchie D. Watson. 1996. Lillian hellman's "the little foxes" and the new south creed: An ironic view of southern history. *The Southern Literary Journal*, 28(2):59–68.

Ikuya Yamada, Ryuji Tamaki, Hiroyuki Shindo, and Yoshiyasu Takefuji. 2018. Studio ousia's quiz bowl question answering system. *arXiv preprint arXiv:1803.08652*.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. In *Proceedings of the International Conference on Learning Representations*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of Empirical Methods in Natural Language Processing*.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. 2018. Generating natural adversarial examples. In *Proceedings of the International Conference on Learning Representations*.

Sentence	Success/Failure Phenomena
its types include "frictional," "cyclical," and "structural" its types include "frictional," and structural	Missing Information ✗
german author of the sorrows of young werther and a two-part faust german author of the sorrows of mr. werther	Lost Named Entity ✗
name this elegy on the death of john keats composed by percy shelley name was this elegy on the death of percy shelley	Incorrect Clue ✗
identify this play about willy loman written by arthur miller so you can identify this work of mr. miller	Unsuited Syntax Template ✗
he employed marco polo and his father as ambassadors he hired marco polo and his father as ambassadors	Verb Synonym ✓

Table 6: Failure and success cases for SCPN. For a majority of Quizbowl questions (97%), the model fails to create a valid paraphrase of the sentence.

A Failure of Syntactically Controlled Paraphrase Networks

We apply the Syntactically Controlled Paraphrase Network (Iyyer et al., 2018, SCPN) to Quizbowl questions. The model operates on the sentence level and cannot paraphrase paragraphs. We thus feed in each sentence independently, ignoring possible breaks in coreference. The model does not correctly paraphrase most of the complex sentences present in Quizbowl questions. The paraphrases were rife with issues: ungrammatical, repetitive, or missing information.

To simplify the setting, we focus on paraphrasing the shortest sentence from each question (often the final clue). The model still fails in this case. We analyze a random sample of 200 paraphrases: only six maintained all of the original information.

Table 6 shows common failure cases. One recurring issue is an inability to maintain the correct named entities after paraphrasing. In Quizbowl, maintaining entity information is vital for ensuring question validity. We were surprised by this failure because SCPN incorporates a copy mechanism.

B Studio Ousia Quizbowl Model

The Studio Ousia system works by aggregating scores from both a neural text classification model and an IR system. Additionally, it scores answers based on their match with the correct entity type (e.g., religious leader, government agency, etc.) predicted by a neural entity type classifier. The Studio Ousia system also uses data beyond Quizbowl questions and the text of Wikipedia pages, integrating entities from a knowledge graph and customized word vectors (Yamada et al., 2018).

C Failed Adversarial Attempt

Figure 10 shows a user's failed attempt to break the neural Quizbowl model.

In his speeches this . . . As a Senator, ~~this man supported **Paraguay** in the **Chaco War**, believing **Bolivia** was backed by Standard Oil.~~ this man's campaign was endorsed by Milo Reno and Charles Coughlin.
Prediction: Huey Long (✓) → Huey Long (✓)

In his speeches this . . . As a Senator, this man's campaign was endorsed by **Milo Reno** and ~~**Charles Coughlin**~~. a Catholic priest and radio show host.
Prediction: Huey Long (✓) → Huey Long (✓)

Figure 10: A failed attempt to trick the neural model. The author modifies the question multiple times, replacing words suggested by the interpretation, but is unable to break the system.