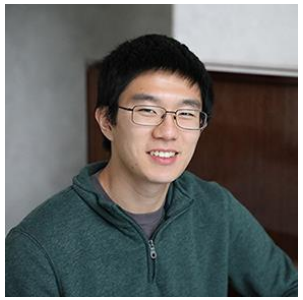
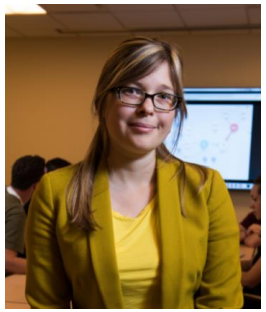
Evaluating Neural Model Robustness for Machine Comprehension





Winston Wu[1] Dustin Arendt[2] Svitlana Volkova[2]

[1]Johns Hopkins University

[2]Pacific Northwest National Laboratory

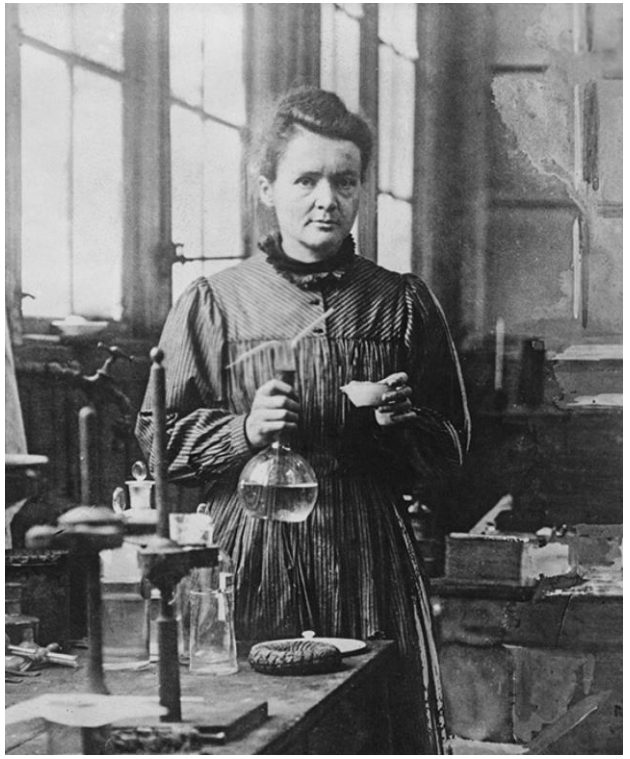
JOHNS HOPKINS UNIVERSITY

CENTER FOR LANGUAGE AND SPEECH PROCESSING

Pacific Northwest NATIONAL LABORATORY

What is Machine Comprehension?

- Context: One of the most famous people born in Warsaw was Maria Skłodowska-Curie, who achieved international recognition for her research on radioactivity and was the first female recipient of the Nobel Prize.



- Question: What was Maria Curie the first female recipient of?

- Answer: Nobel Prize

Research Questions

- How robust are MC models to different types and amounts of perturbations?
- What factors of the data contribute to model errors?

Data

- SQuAD [Rajpurkar+ 2016]
- TriviaQA [Joshi+ 2017]

Models

- BiDAF w/ ELMo [Seo+ 2017, Peters+ 2018]
- BERT [Devlin+ 2019]

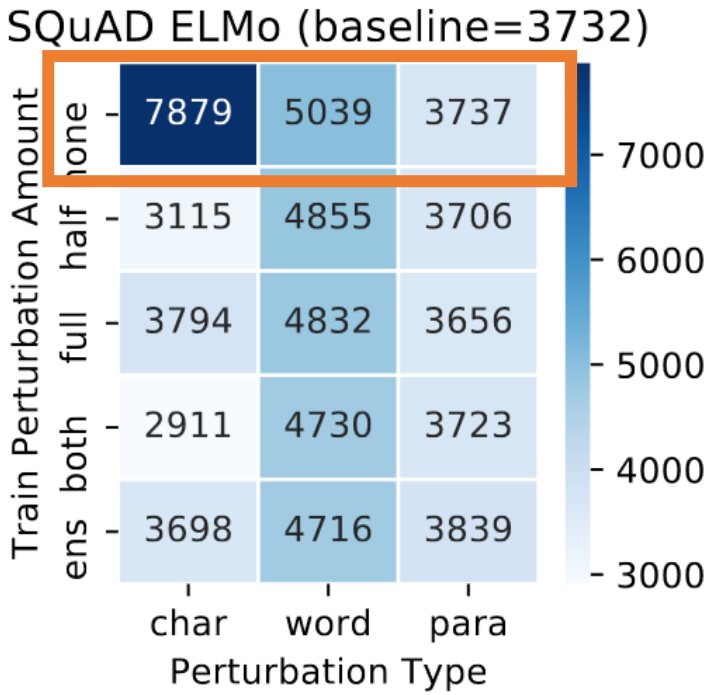
Perturbations

Original	The connection between macroscopic nonconservative forces and microscopic conservative forces is described by detailed treatment with statistical mechanics.
Character Replacement • Identical looking but with different Unicode codepoints • Homograph attack	The connection between macroscopic nonconservative forces and microscopic conservative forces is described by detailed treatment with statistical mechanics.
Word Replacement • Replace with nearest neighbor	The connection between macroscopic nonconservative forces and insect conservative troops is referred by detailed treatment with statistical mechanics.
Sentence Paraphrase • Paraphrase with Improved ParaBank Rewriter [Hu+ 2019]	The link between macroscopic non-conservative forces and microscopic conservative forces is described in detail by statistical mechanics.

Experiments

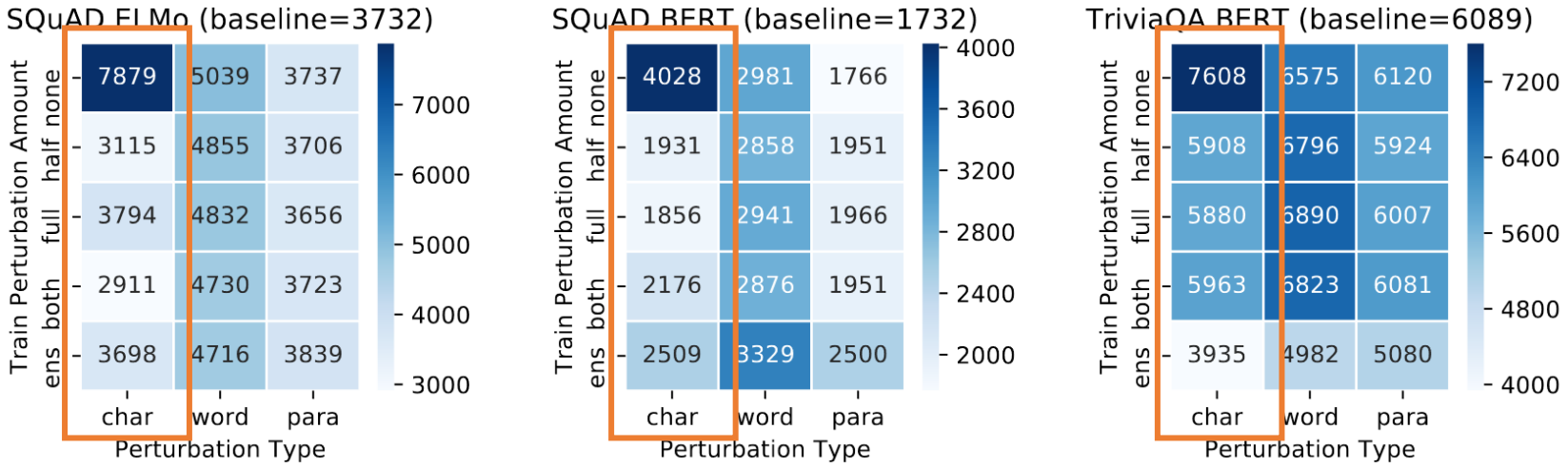
Clean training data

Adversarial training



Higher (darker)
= more errors

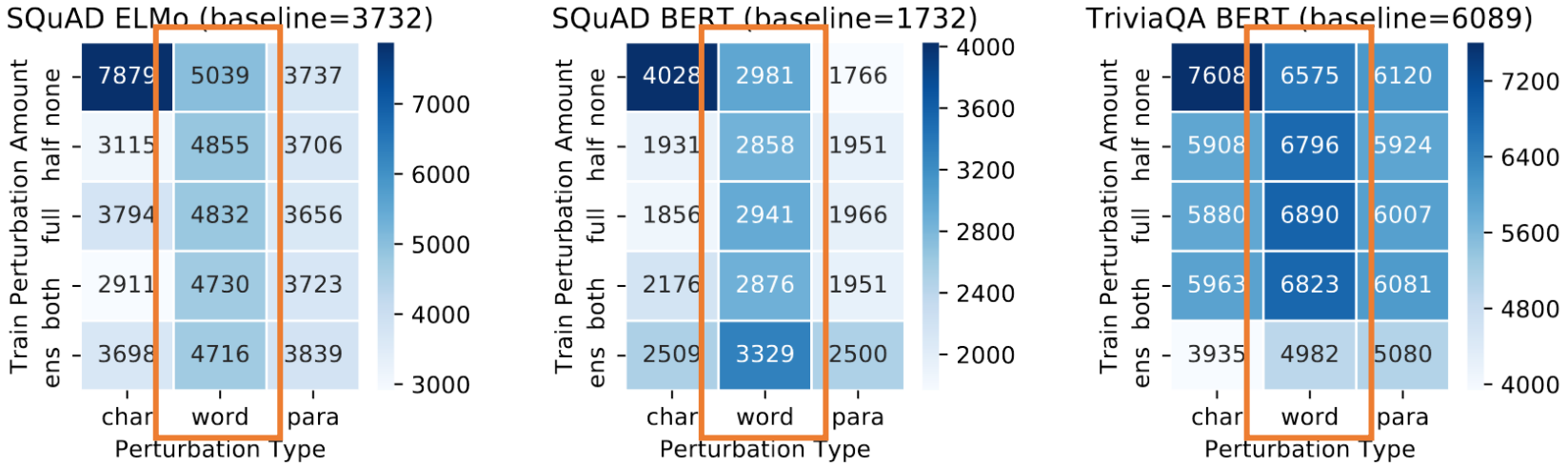
Effects of Character Perturbations



(a) SQuAD, ELMo (b) SQuAD, BERT (c) TriviaQA, BERT

- Character perturbations are the most harmful
- But are the most easily made robust against

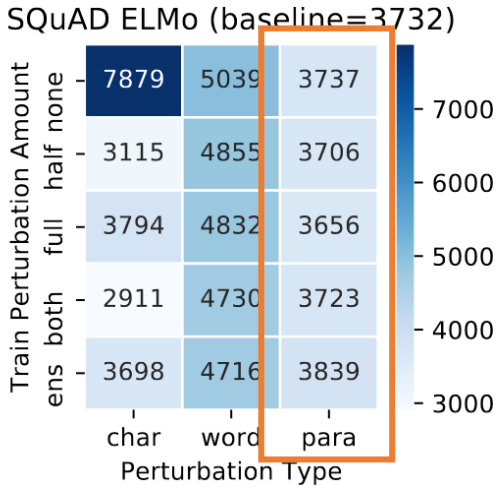
Effects of Word Perturbations



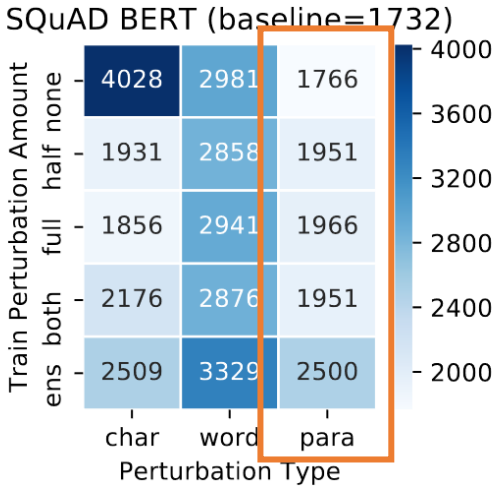
(a) SQuAD, ELMo (b) SQuAD, BERT (c) TriviaQA, BERT

- Word perturbations introduce modest amount of errors
- Adversarial training does not seem to help

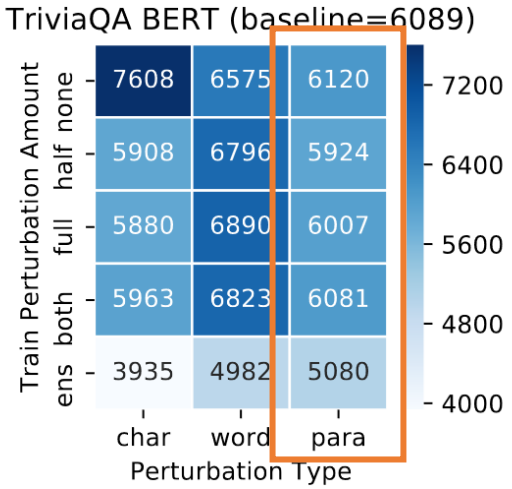
Effects of Sentence Perturbations



(a) SQuAD, ELMo



(b) SQuAD, BERT



(c) TriviaQA, BERT

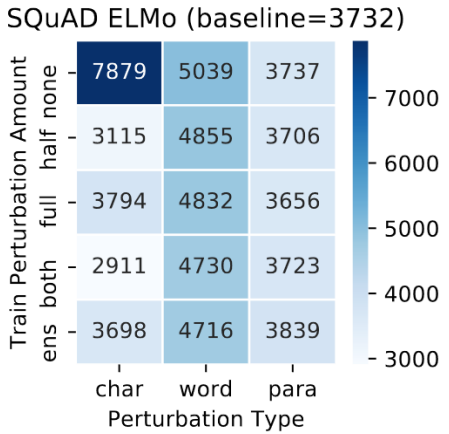
- Introduces the least errors
- Improve with strategic paraphrasing

Strategic Paraphrasing

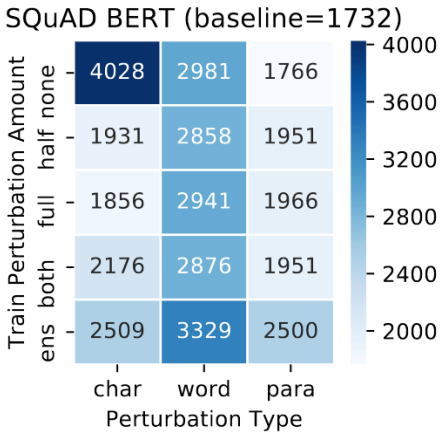
- Identify important words by removing a word and checking the model's prediction and confidence
- Rewrite the most important words using constrained decoding

Original Paragraph	Strategic Paraphrase
. . . Veteran receiver Demaryius Thomas led the team with 105 receptions for 1,304 yards and six touchdowns, while Emmanuel Sanders caught 76 passes for 1,135 yards and six scores, while adding another 106 yards returning punts.	. . . The veteran earman Demaryius Thomas was leading a team of 1,304 yards and six touchdowns, while Emmanuel Sanders caught 76 passes for 1,135 yards and six scores while he added another 106 yards of punts back.
Question: Who led the Broncos with 105 receptions? **Answer**: Demaryius Thomas (correct) → Emmanuel Sanders (incorrect)	

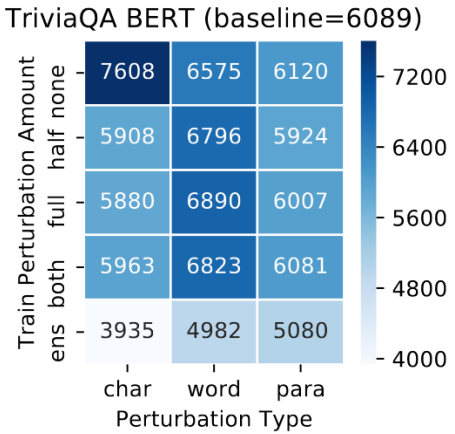
General Observations



(a) SQuAD, ELMo



(b) SQuAD, BERT



(c) TriviaQA, BERT

- When training with perturbed data, the amount of perturbed data does not matter too much
- BERT model made less errors than ELMo
- TriviaQA may be a harder dataset than SQuAD
- Ensembling helped for TriviaQA

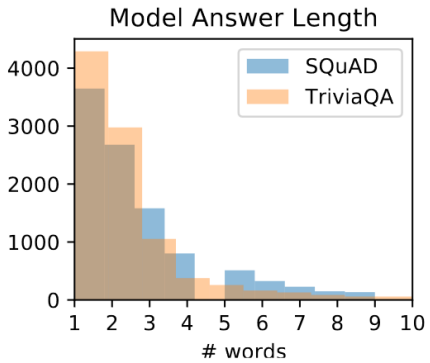
Explaining Model Performance Through Data

- What factors of the data contribute to model errors?
  - Model answer length
  - Question type
  - Question complexity
  - Context complexity
- CrossCheck [Arendt+ 2020]
- Predicting model errors

(d) Error distribution in BERT models across different perturbations and question types

Factors Associated With Errors

- Model answer length
  - Longer answers returned by the model are more likely to be incorrect

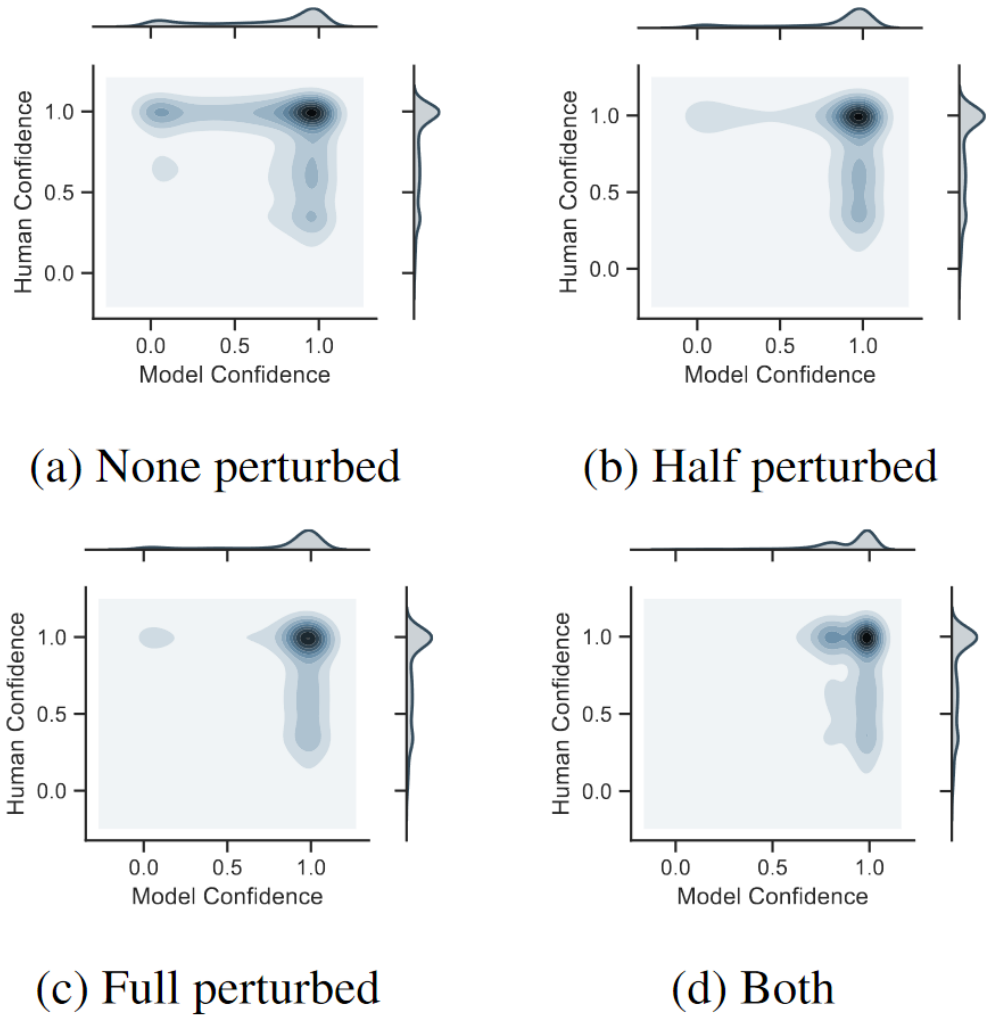


- Question Type
  - Some questions are easier to answer
    - “When” and “How many”
    - Even when incorrect, answers tend to be the right type

Factors Associated With Errors

- Question Difficulty: inter-annotator agreement (SQuAD)
  - Low-confidence when trained on clean data
  - More confident when adversarially trained
- Context Difficulty: Flesch-Kincaid readability

Data	Correct	Errors
SQuAD	12.9	13.0
TriviaQA	17.1	17.5

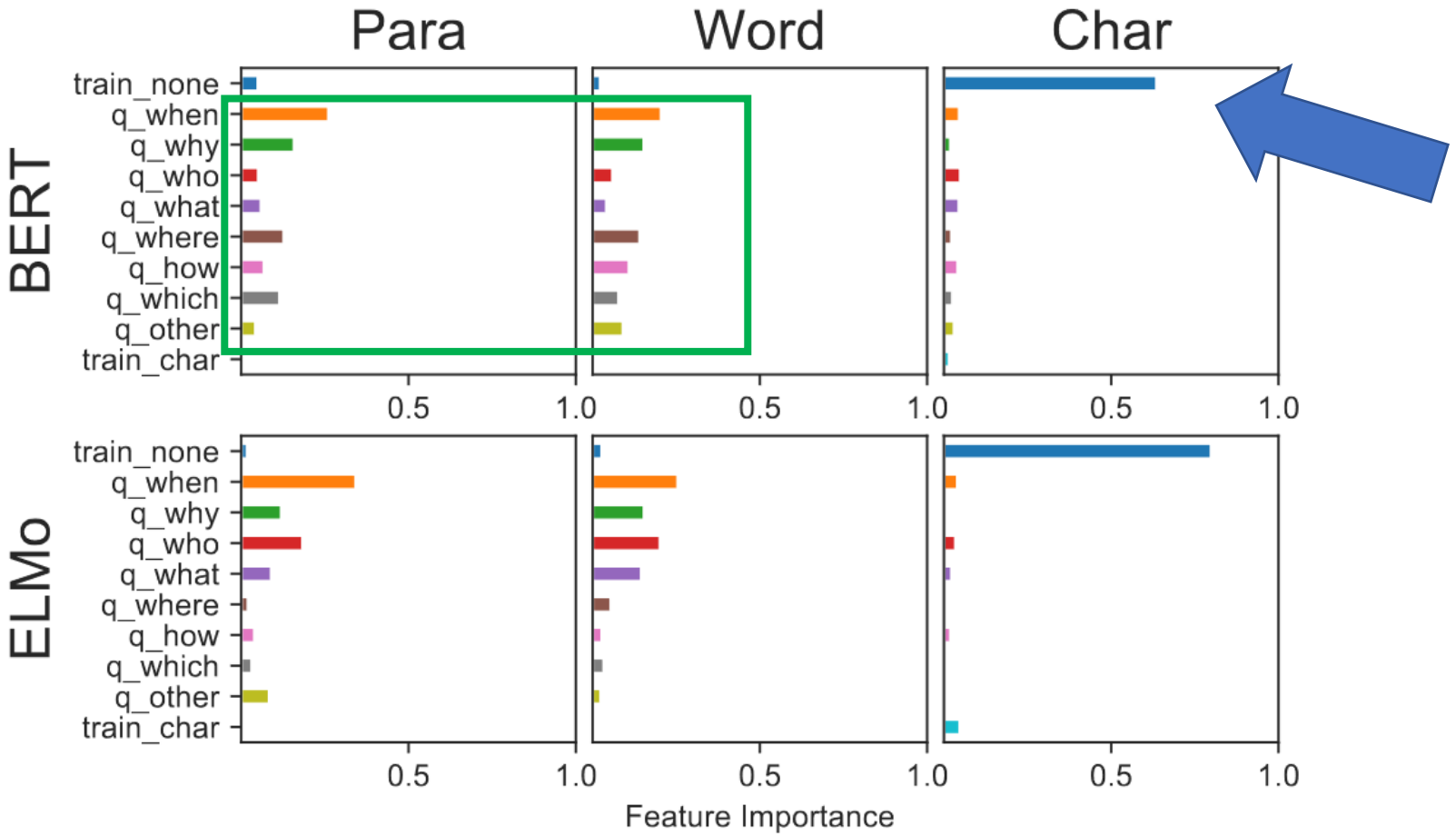


(a) None perturbed (b) Half perturbed

(c) Full perturbed (d) Both

Predicting Model Errors

- Binary classification using XGBoost



Summary

- How robust are MC models to different types and amounts of perturbations?
  - BERT is more robust then BiDAF+ELMo
  - Adversarial training helps
  - Strategic paraphrasing: adversarially rewrite important words
- What factors of the data contribute to model errors?
  - Training amount, perturbation type, question type, question length, context length, answer length, context and question complexity
  - Created a model to predict errors

- I'm looking for a postdoc position! Check out my other work:

cs.jhu.edu/~winston