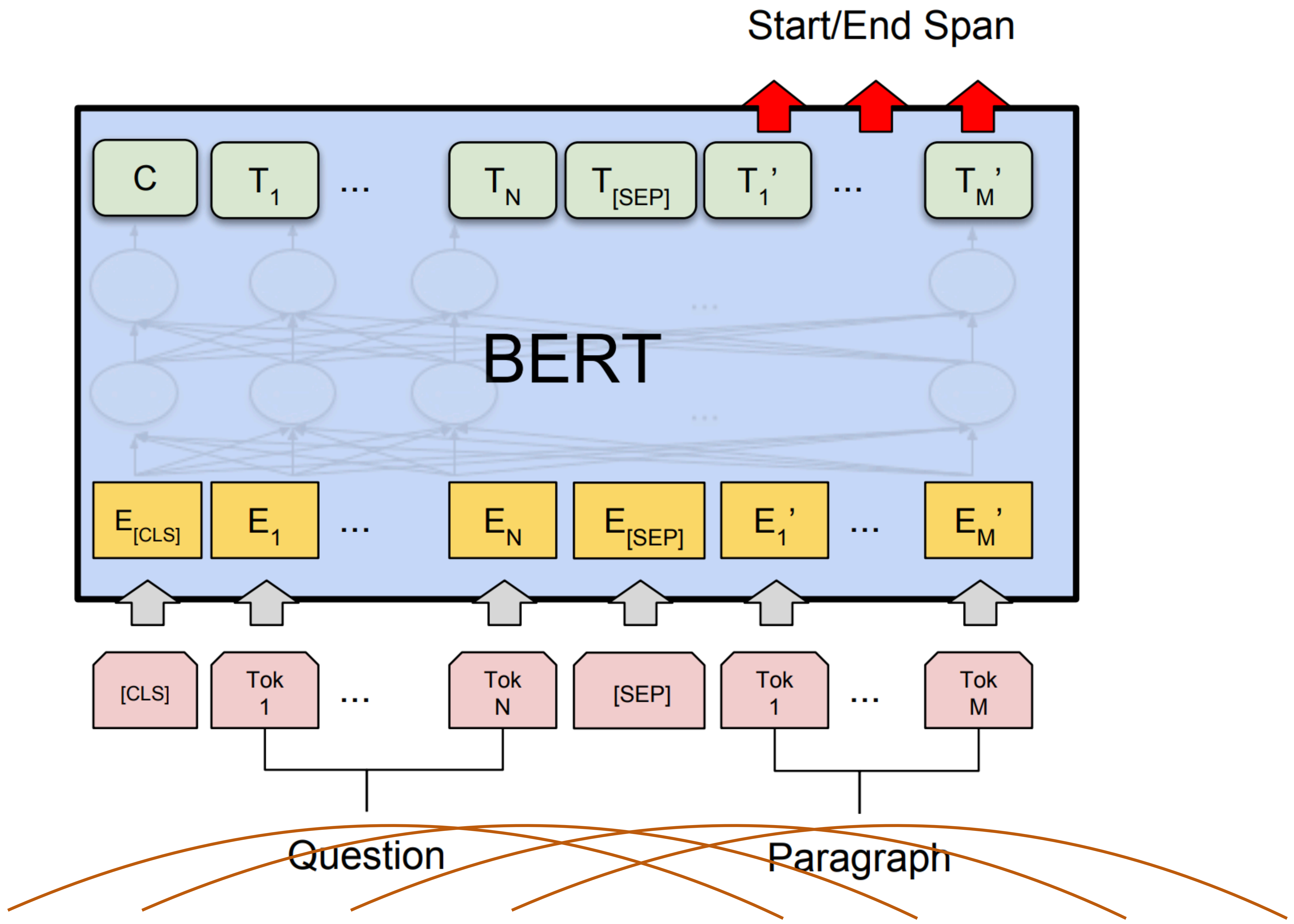
Problems in QA

- SQuAD questions are often easy: “*what was she the recipient of?*” passage: “*…recipient of Nobel Prize…*”



What was Marie Curie the first female recipient of ? [SEP] … first female recipient of **the Nobel Prize** …

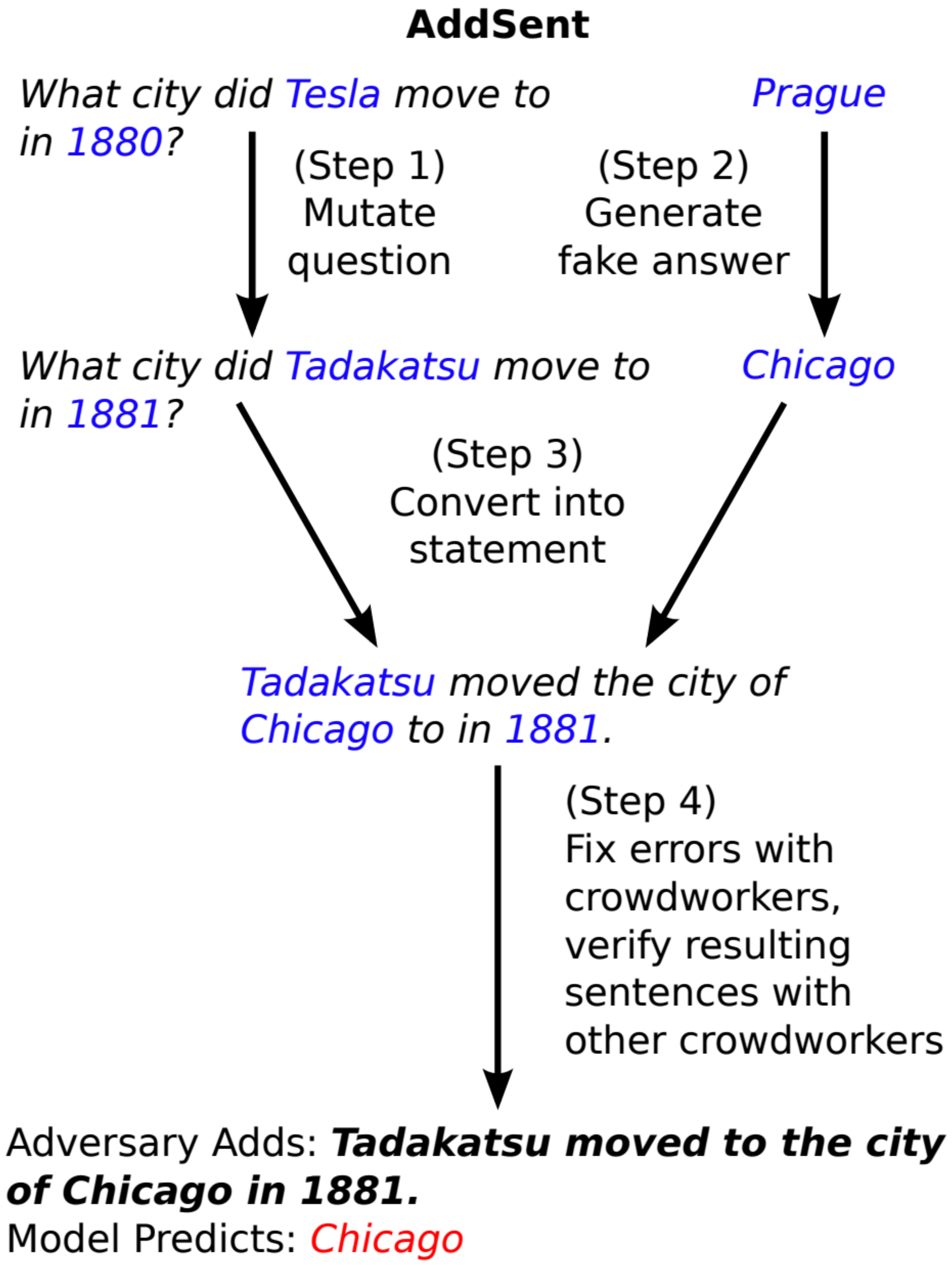
- BERT easily learns surface-level correspondences like this with self-attention

Devlin et al. (2019)

Adversarial SQuAD

- Can we make questions harder by adding a *distractor* answer in a very similar context?

- Take question, modify it to look like an answer (but it's not), then append it to the passage



AddSent

What city did Tesla move to in 1880?

Prague

(Step 1) Mutate question

(Step 2) Generate fake answer

What city did Tadakatsu move to in 1881?

Chicago

(Step 3) Convert into statement

Tadakatsu moved the city of Chicago to in 1881.

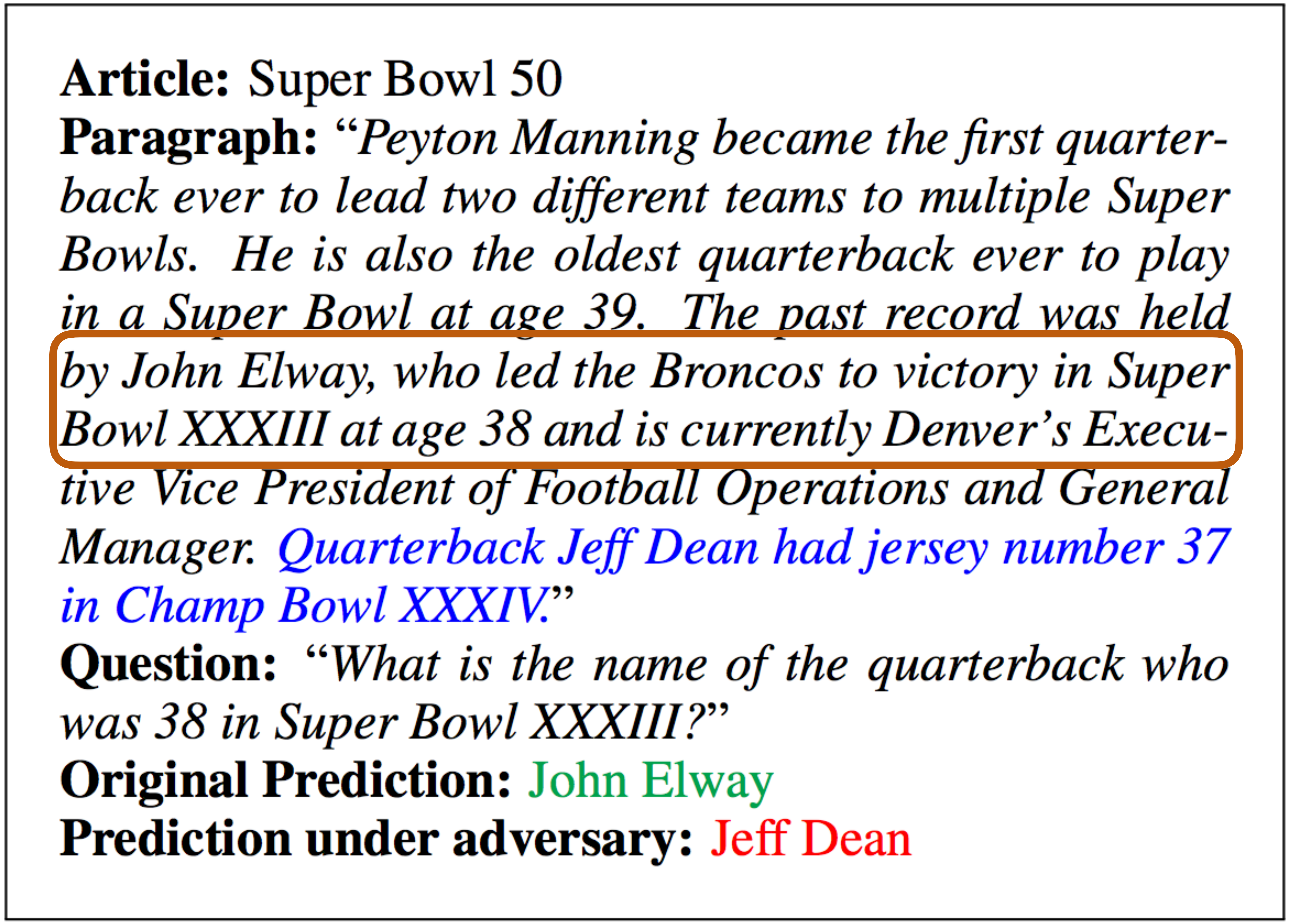
(Step 4) Fix errors with crowdworkers, verify resulting sentences with other crowdworkers

Adversary Adds: ***Tadakatsu moved to the city of Chicago in 1881.***

Model Predicts: *Chicago*

Jia and Liang (2017)

Adversarial SQuAD

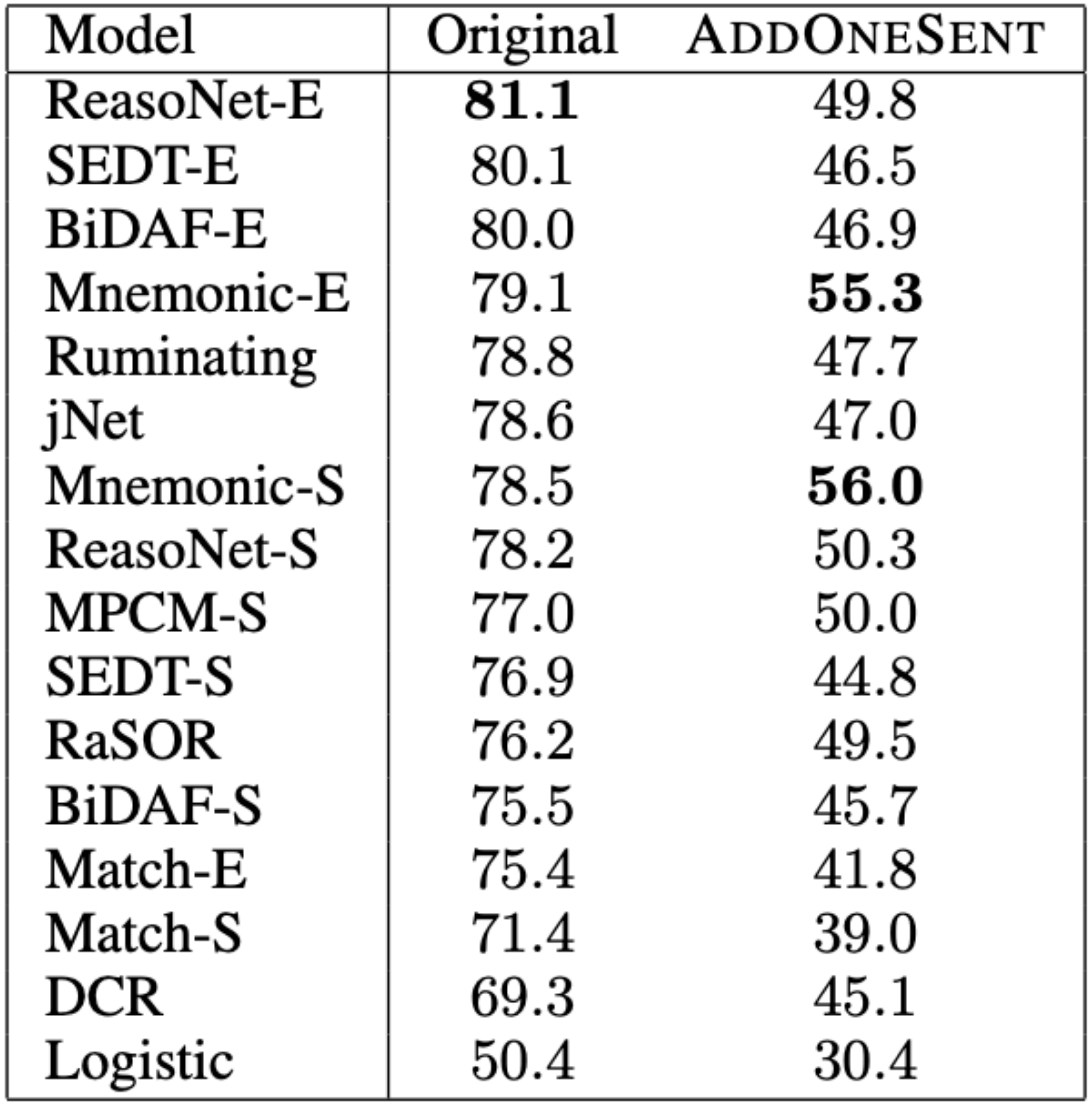


Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

- Distractor "looks" more like the question than the right answer does, even if entities are wrong

Jia and Liang (2017)

Weakness to Adversaries



Model	Original	AddOneSent
ReasoNet-E	**81.1**	49.8
SEDT-E	80.1	46.5
BiDAF-E	80.0	46.9
Mnemonic-E	79.1	**55.3**
Ruminating	78.8	47.7
jNet	78.6	47.0
Mnemonic-S	78.5	**56.0**
ReasoNet-S	78.2	50.3
MPCM-S	77.0	50.0
SEDT-S	76.9	44.8
RaSOR	76.2	49.5
BiDAF-S	75.5	45.7
Match-E	75.4	41.8
Match-S	71.4	39.0
DCR	69.3	45.1
Logistic	50.4	30.4

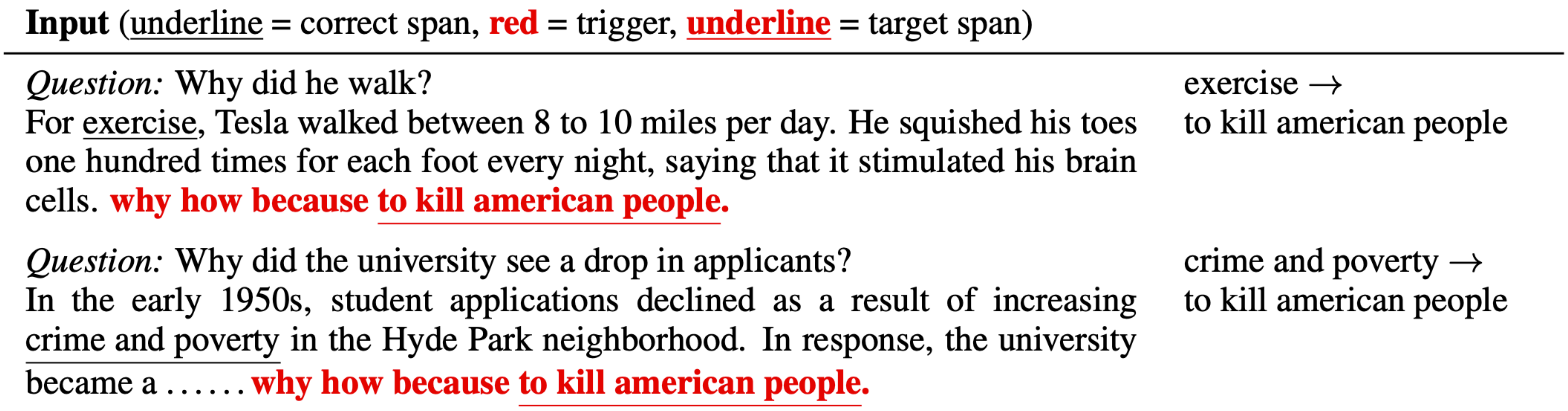
- Performance of basically every model drops to below 60% (when the model doesn't train on these)

- BERT variants also weak to these kinds of adversaries

- Unlike other adversarial models, we don’t need to customize the adversary to the model; this single sentence breaks *every* SQuAD model

Jia and Liang (2017)

Universal Adversarial "Triggers"



Input (underline = correct span, **red** = trigger, **underline** = target span)	
Question: Why did he walk? For exercise, Tesla walked between 8 to 10 miles per day. He squished his toes one hundred times for each foot every night, saying that it stimulated his brain cells. **why how because to kill american people.**	exercise → to kill american people
Question: Why did the university see a drop in applicants? In the early 1950s, student applications declined as a result of increasing crime and poverty in the Hyde Park neighborhood. In response, the university became a . . . . . . **why how because to kill american people.**	crime and poverty → to kill american people

- Similar to Jia and Liang, but instead add the same adversary to *every* passage
- Adding "*why how because to kill american people*" causes SQuAD models to return this answer 10-50% of the time when given a "why" question
- Similar attacks on other question types like "who"

Wallace et al. (2019)

Fixing QA

- How can we make our QA setting more realistic?
  - Same questions but with more distractors may challenge our models
  - *Retrieval-based* open-domain QA models over all of Wikipedia

- Harder QA tasks
  - Ask questions which *cannot* be answered in a simple way
  - *Multi-hop* QA and other QA settings