Can NLI Models Verify QA Systems' Predictions?

Jifan Chen, Eunsol Choi and Greg Durrett

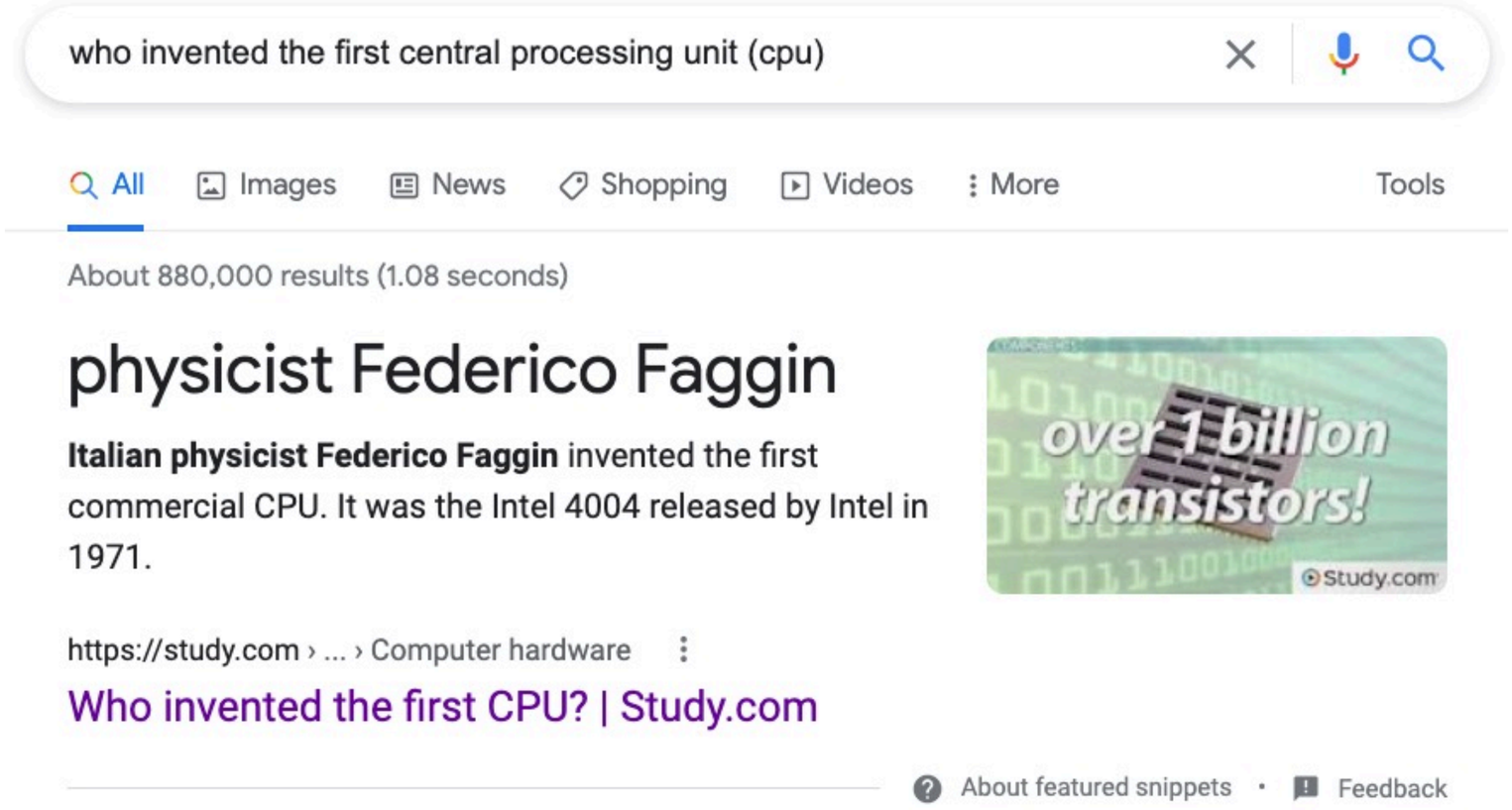
The University of Texas at Austin

Unreliable Predictions of QA systems

- Current QA systems try to return a most-plausible answer to users:

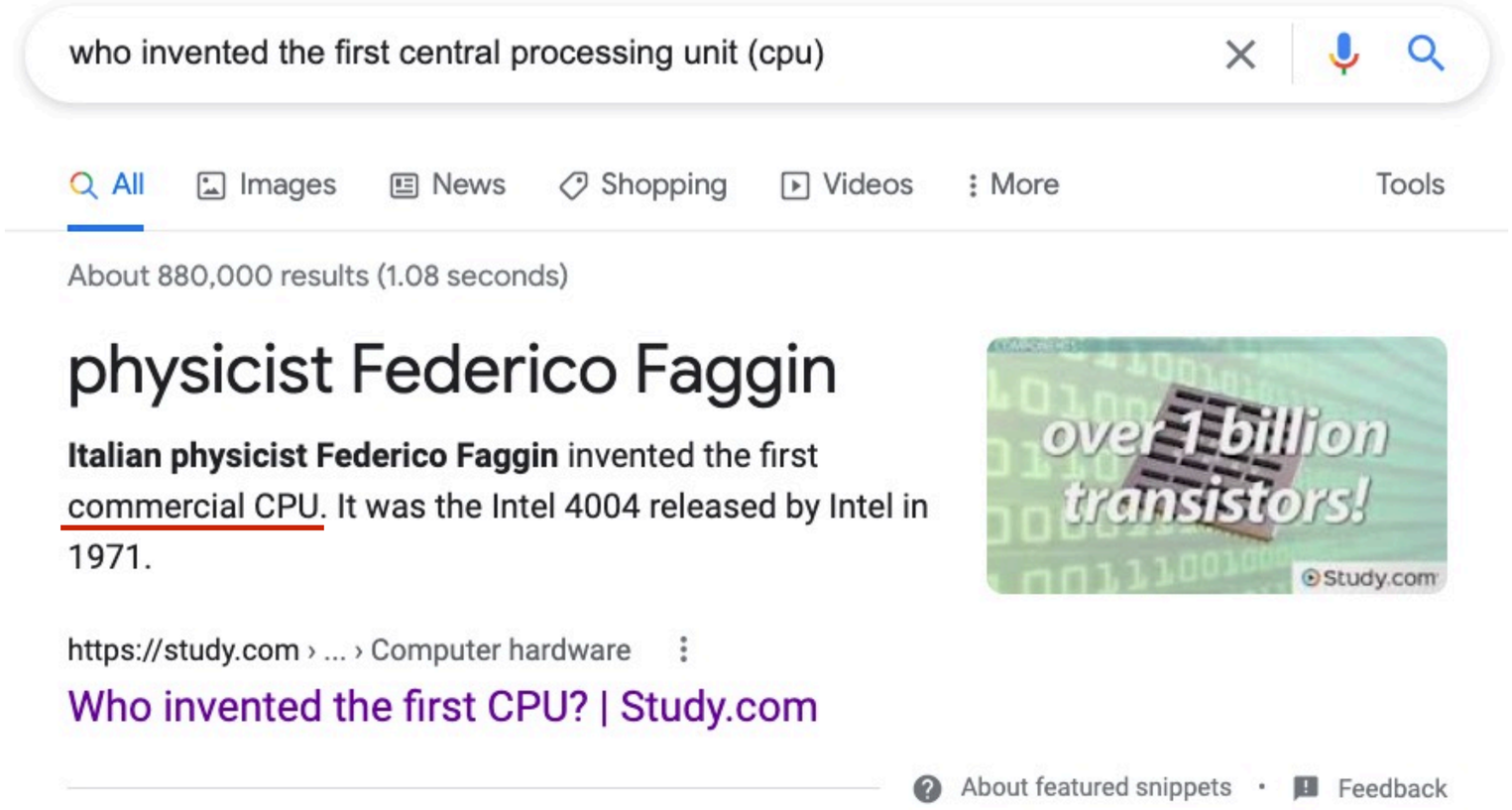
Unreliable Predictions of QA systems

- Current QA systems try to return a most-plausible answer to users:



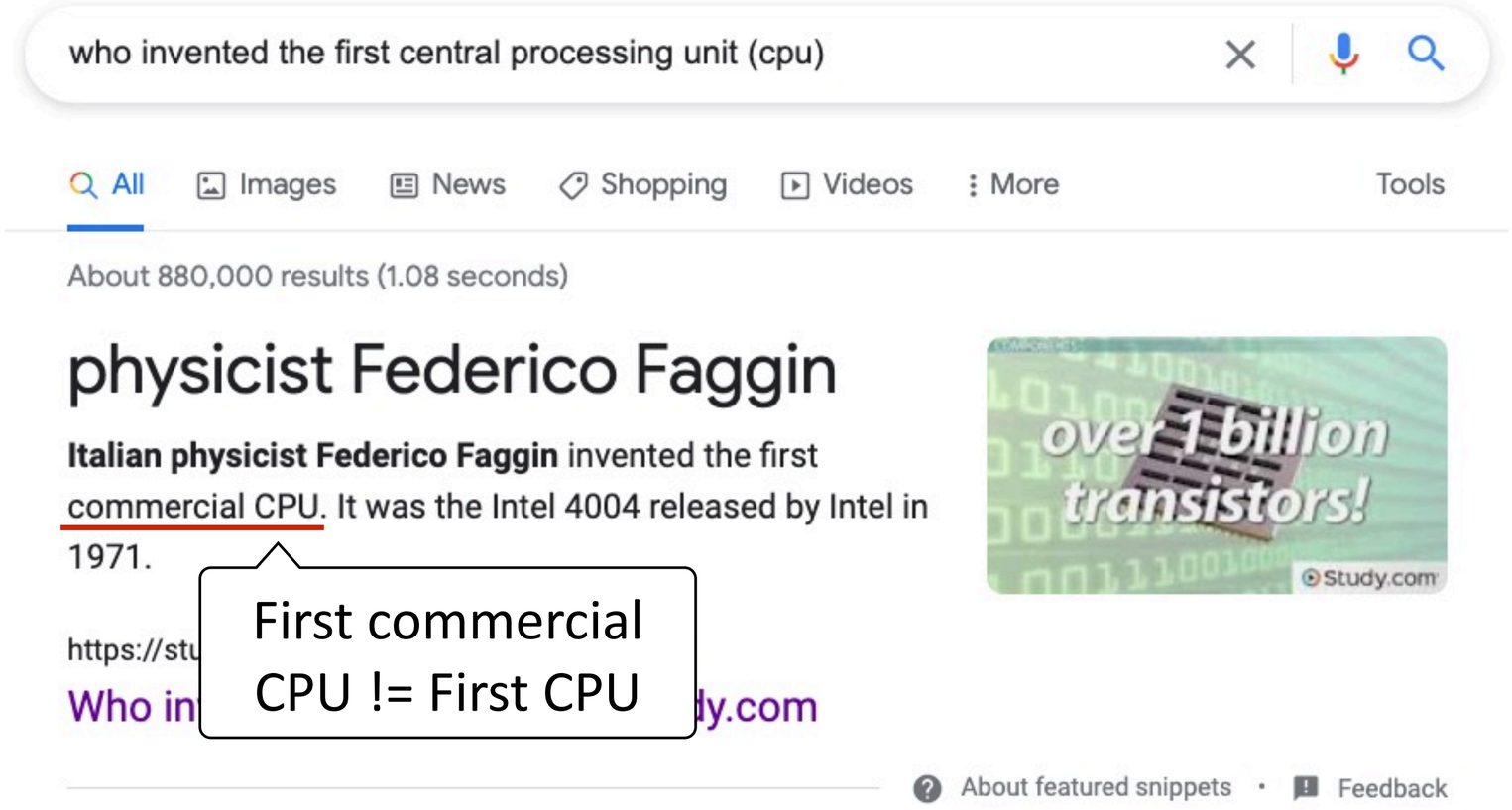
Unreliable Predictions of QA systems

- Current QA systems try to return a most-plausible answer to users:



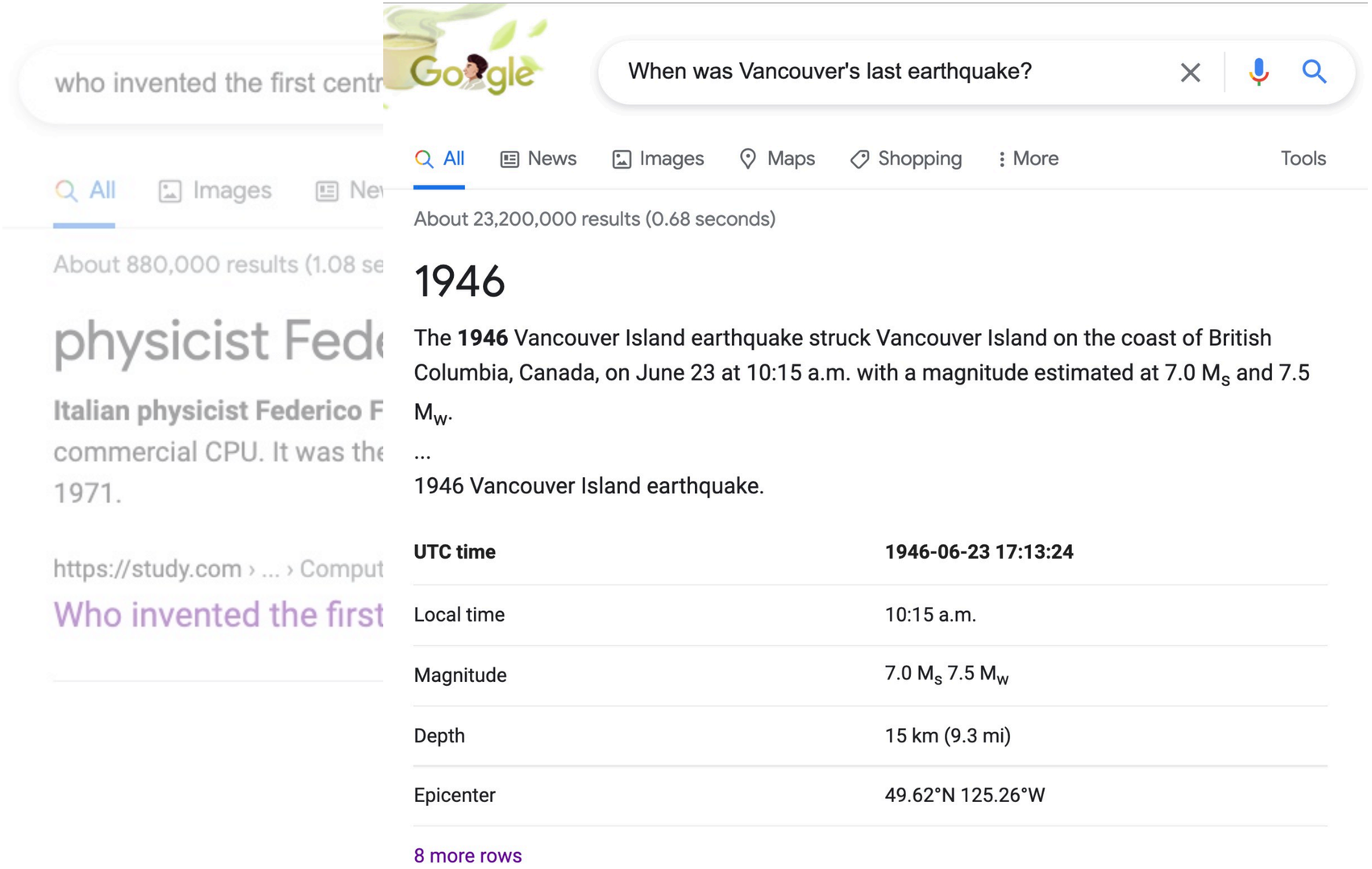
Unreliable Predictions of QA systems

- Current QA systems try to return a most-plausible answer to users:



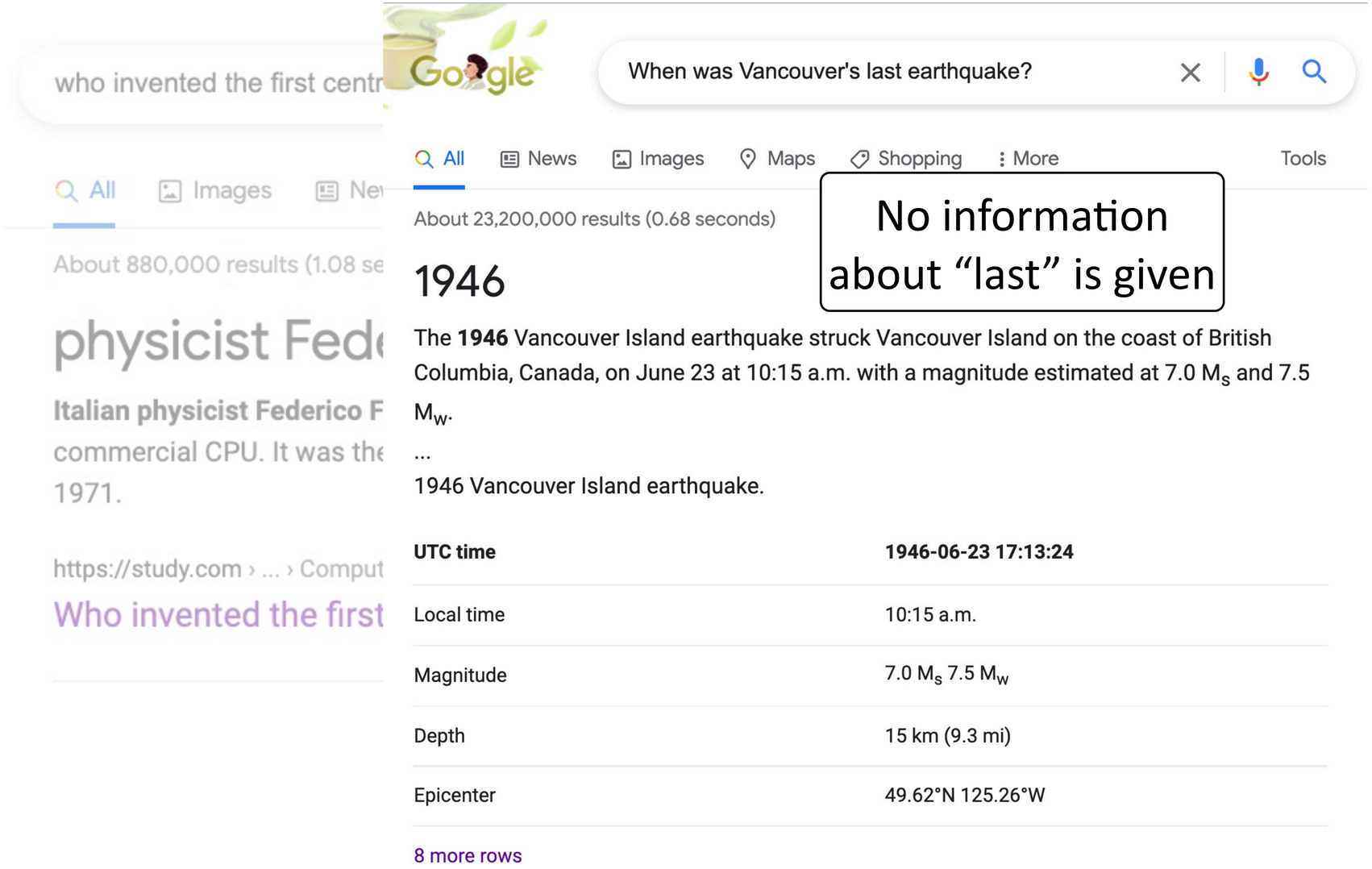
Unreliable Predictions of QA systems

- Current QA systems try to return a most-plausible answer to users:



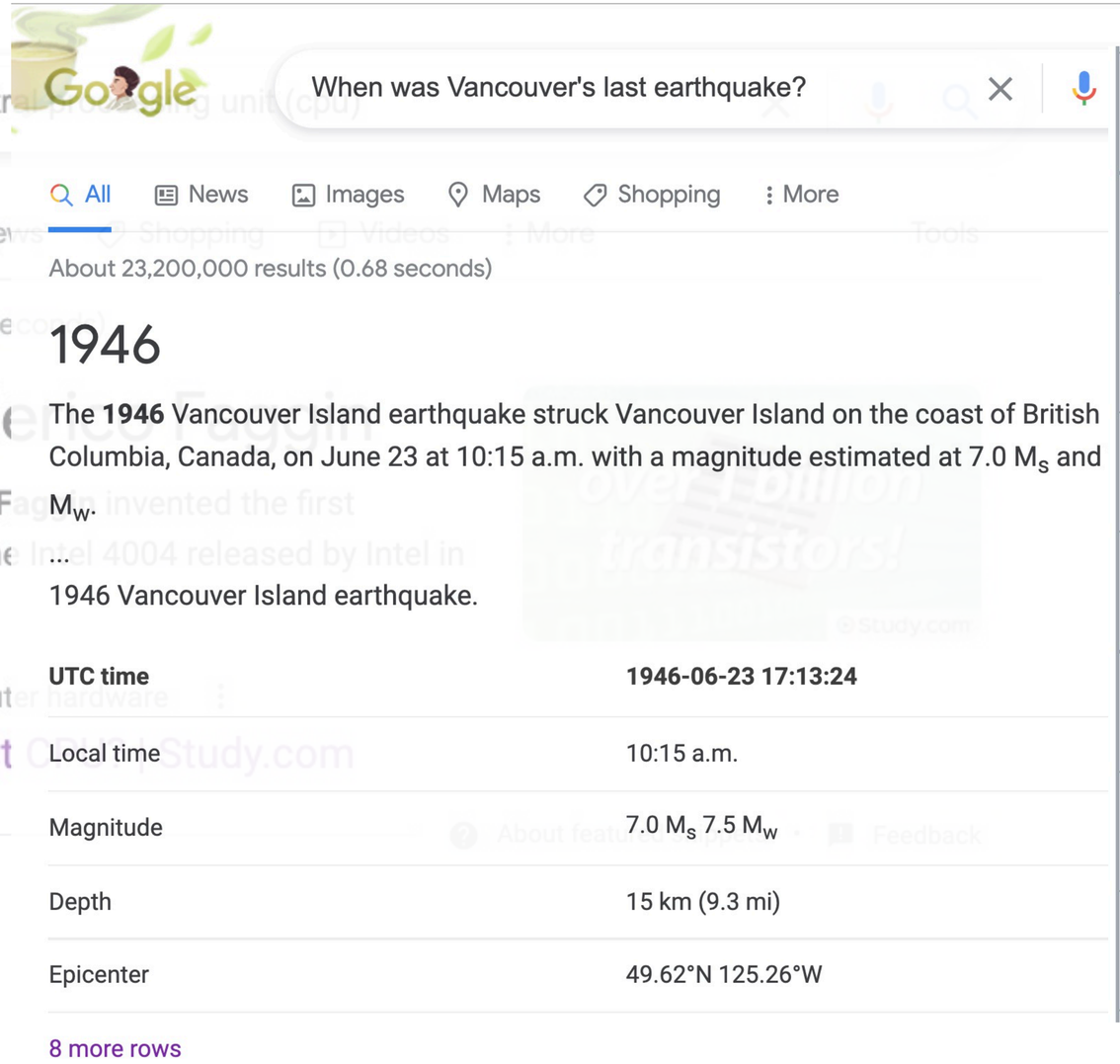
Unreliable Predictions of QA systems

- Current QA systems try to return a most-plausible answer to users:



Unreliable Predictions of QA systems

- Current QA systems try to return a most-plausible answer to users:



2011-09-09	Vancouver Island
2010-06-23	Central Canada
2009-11-17	Queen Charlotte Islands, BC
2009-07-07	Baffin Bay
2008-01-05	Queen Charlotte Islands, BC
2007-10-09	The Nazko region
2004-11-02	Vancouver Island, BC
2004-07-19	Vancouver Island

NLI as a verifier

- Idea: Use Natural Language Inference (NLI) to verify whether an answer can be **truly** entailed from its corresponding context (Harabagiu and Hickl, 2006; Peñas et al.2008; Yin et al. 2021; Mishra et al. 2021).

Example picked from Natural Questions (Kwiatkowski et al. 2019)

NLI as a verifier

- Idea: Use Natural Language Inference (NLI) to verify whether an answer can be **truly** entailed from its corresponding context (Harabagiu and Hickl, 2006; Peñas et al.2008; Yin et al. 2021; Mishra et al. 2021).

Question: What is the revolution period of Venus in earth days?
Answer: 243 days

Question conversion

Hypothesis: The **revolution** period of Venus in earth days is 243 days.

Example picked from Natural Questions (Kwiatkowski et al. 2019)

NLI as a verifier

- Idea: Use Natural Language Inference (NLI) to verify whether an answer can be **truly** entailed from its corresponding context (Harabagiu and Hickl, 2006; Peñas et al.2008; Yin et al. 2021; Mishra et al. 2021).

Question: What is the revolution period of Venus in earth days?
Answer: 243 days

Question conversion

Hypothesis: The **revolution** period of Venus in earth days is 243 days.

Context: Venus is the second planet from the Sun… . **It** has the longest rotation period (243 days)…

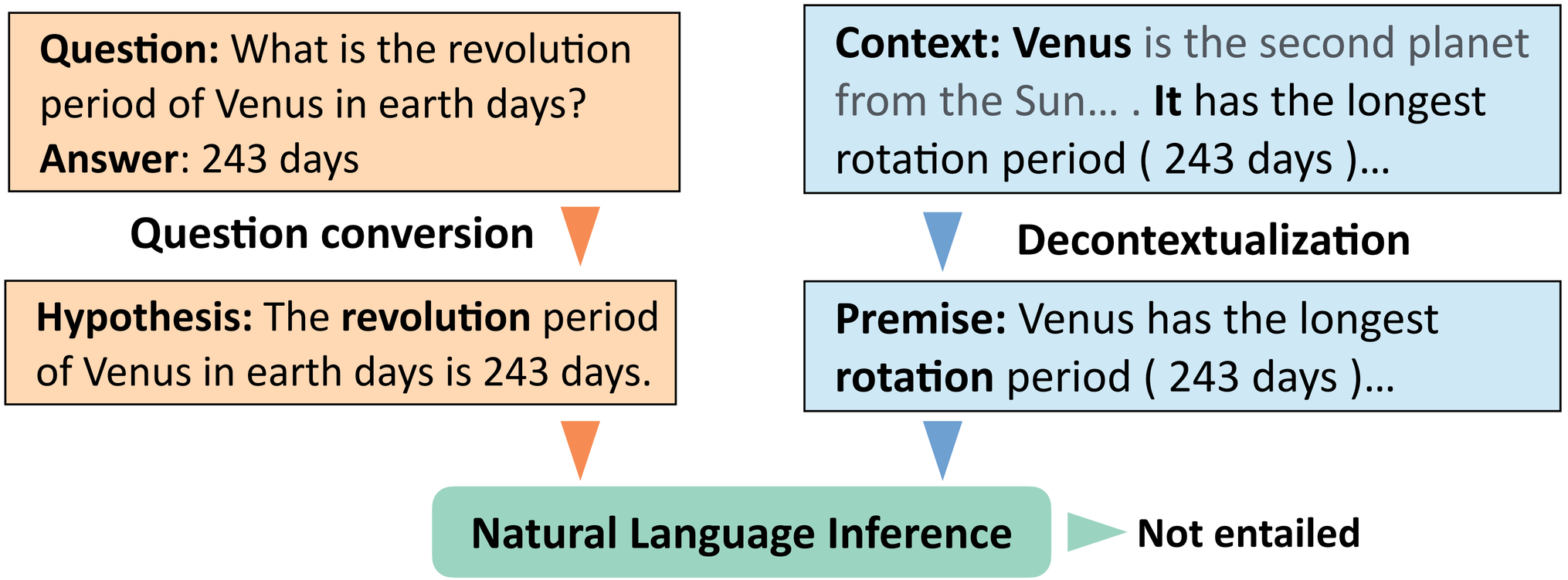
Decontextualization

Premise: Venus has the longest **rotation** period (243 days)…

Example picked from Natural Questions (Kwiatkowski et al. 2019)

NLI as a verifier

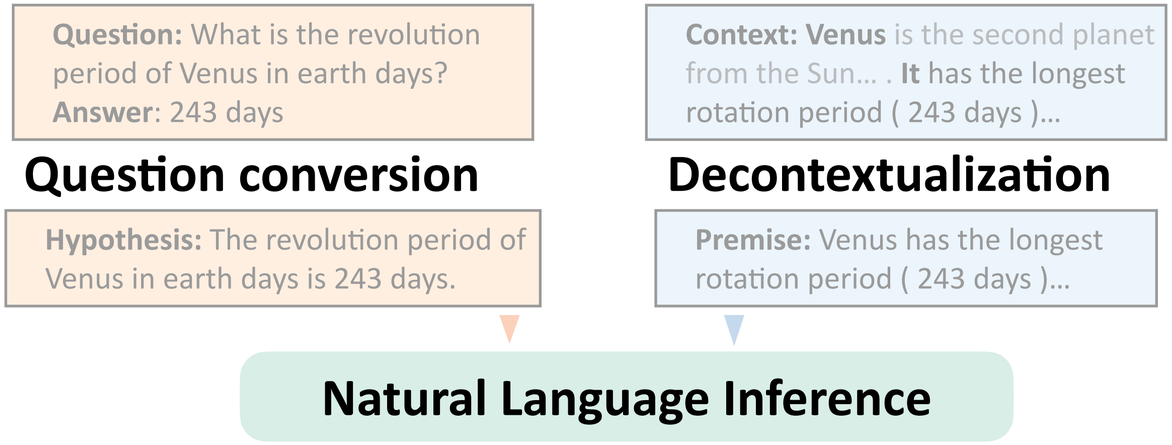
- Idea: Use Natural Language Inference (NLI) to verify whether an answer can be **truly** entailed from its corresponding context (Harabagiu and Hickl, 2006; Peñas et al.2008; Yin et al. 2021; Mishra et al. 2021).



Example picked from Natural Questions (Kwiatkowski et al. 2019)

Outline

1) NLI as a QA Verifier



2) Experiments

- Rejecting unanswerable questions
- Improving the prediction confidence
- Disagreement between NLI and QA

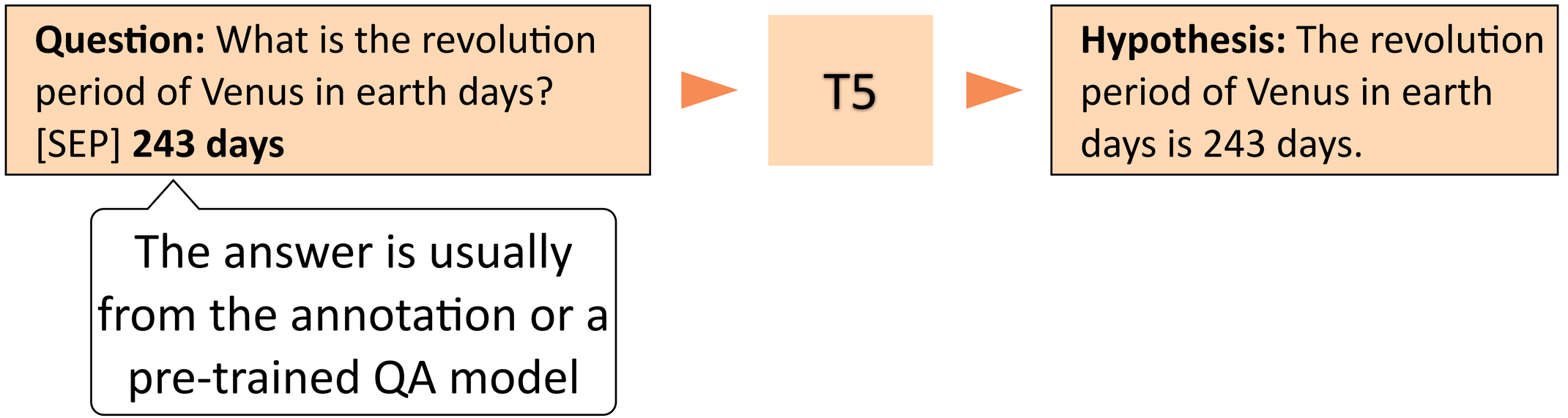
3) Takeaways

Question Conversion

- Demszky et al. (2018) explored a rule-based conversion system.
- This work: A T5-based model converting a (question, answer) pair to a statement trained on the annotations from Demszky et al. (2018).

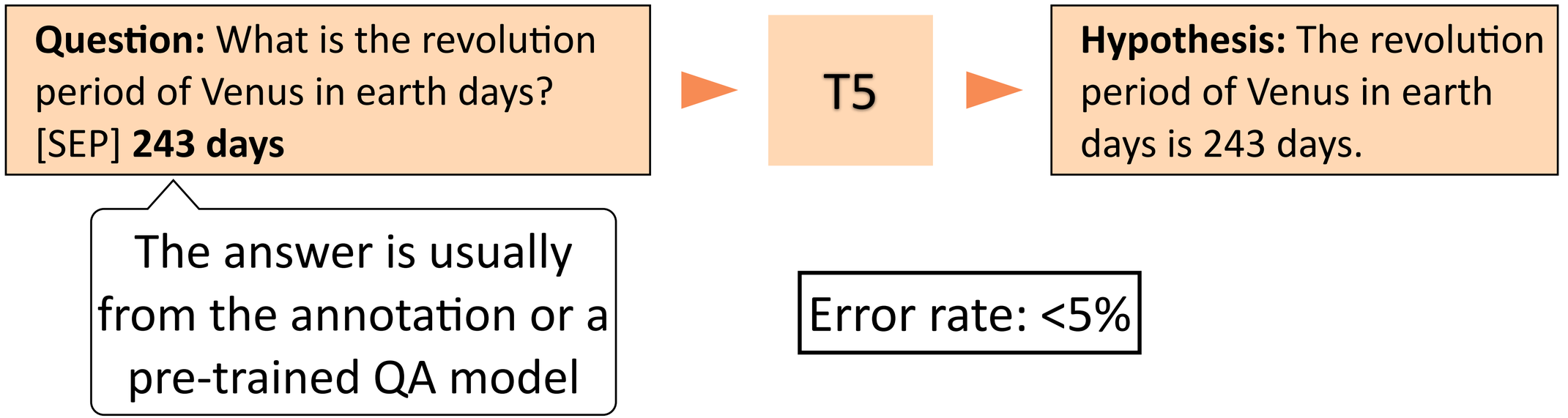
Question Conversion

- Demszky et al. (2018) explored a rule-based conversion system.
- This work: A T5-based model converting a (question, answer) pair to a statement trained on the annotations from Demszky et al. (2018).



Question Conversion

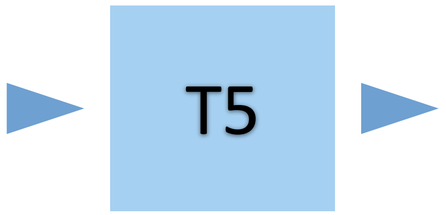
- Demszky et al. (2018) explored a rule-based conversion system.
- This work: A T5-based model converting a (question, answer) pair to a statement trained on the annotations from Demszky et al. (2018).



Decontextualization

- The whole paragraph contains too much information
- A T5-based model to rewrite (name completion, NP/pronoun swap, bridging) a sentence to be interpretable out of context if feasible (Choi et al. 2021).

Venus is the second planet from the Sun , orbiting it every 224.7 Earth days . **It** has the longest rotation period (243 days) of any planet in the Solar System and rotates in the opposite direction to most other planets …



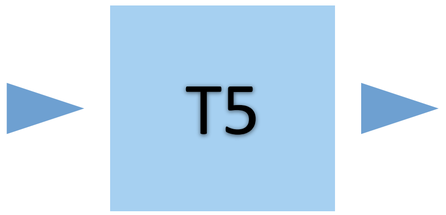
Venus has the longest rotation period (243 days) of any planet in the Solar System and rotates in the opposite direction to most other planets …

Decontextualization

- The whole paragraph contains too much information
- A T5-based model to rewrite (name completion, NP/pronoun swap, bridging) a sentence to be interpretable out of context if feasible (Choi et al. 2021).

Error rate: <10%

Venus is the second planet from the Sun , orbiting it every 224.7 Earth days . **It** has the longest rotation period (243 days) of any planet in the Solar System and rotates in the opposite direction to most other planets …



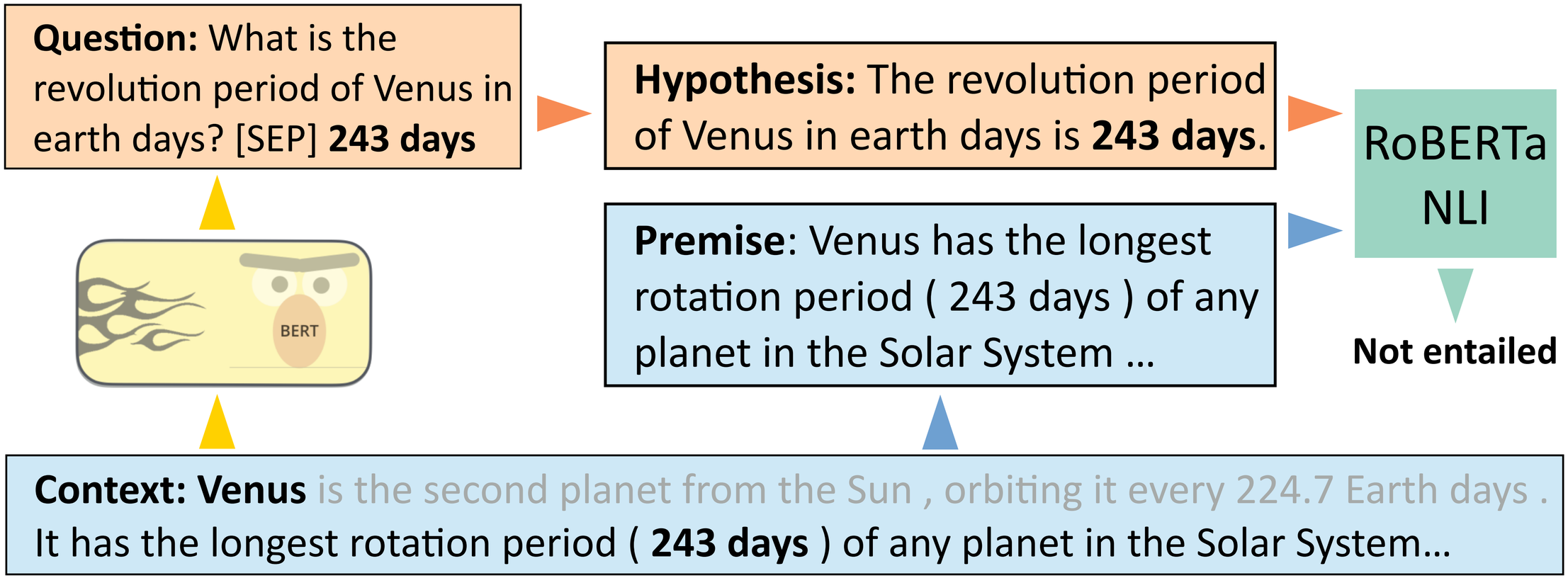
Venus has the longest rotation period (243 days) of any planet in the Solar System and rotates in the opposite direction to most other planets …

NLI model

- A Roberta-based model trained with the (premise, hypothesis) pairs from QA datasets: gold answer + context as positive, non-gold answers in the top-k predictions as negatives

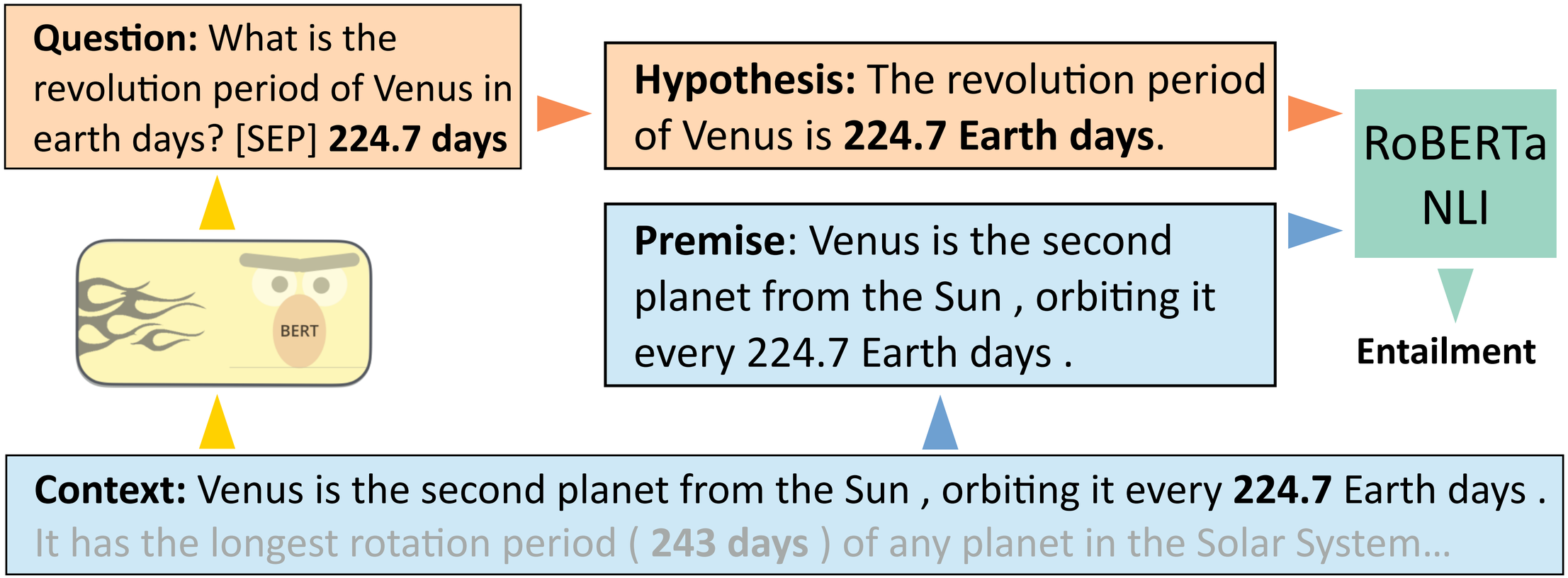
NLI model

- A Roberta-based model trained with the (premise, hypothesis) pairs from QA datasets: gold answer + context as positive, non-gold answers in the top-k predictions as negatives



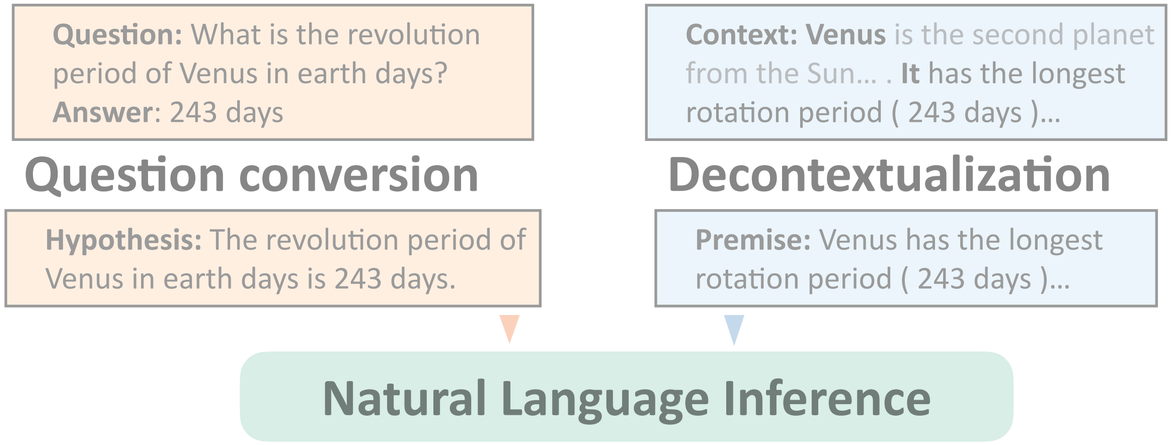
NLI model

- A Roberta-based model trained with the (premise, hypothesis) pairs from QA datasets: gold answer + context as positive, non-gold answers in the top-k predictions as negatives



Outline

1) NLI as a QA Verifier



2) Experiments

- Rejecting unanswerable questions
- Improving the prediction confidence
- Disagreement between NLI and QA

3) Takeaways

Rejecting unanswerable questions

Rejecting unanswerable questions

- Experimental setup:
  - Train a QA model on SQuAD1.1 (every question is answerable) and test it on SQuAD2.0 (contains unanswerable questions).
  - Using an NLI model pre-trained on MNLI to verify the predictions from the SQuAD1.1 model (always gives an answer).

Rejecting unanswerable questions

- Experimental setup:
  - Train a QA model on SQuAD1.1 (every question is answerable) and test it on SQuAD2.0 (contains unanswerable questions).
  - Using an NLI model pre-trained on MNLI to verify the predictions from the SQuAD1.1 model (always gives an answer).

- Results: MNLI model successfully rejects **78.5%** of the unanswerable examples and accepts **82.5%** of the answerable examples.

Rejecting unanswerable questions

- Experimental setup:
  - Train a QA model on SQuAD1.1 (every question is answerable) and test it on SQuAD2.0 (contains unanswerable questions).
  - Using an NLI model pre-trained on MNLI to verify the predictions from the SQuAD1.1 model (always gives an answer).

- Results: MNLI model successfully rejects **78.5%** of the unanswerable examples and accepts **82.5%** of the answerable examples.
  - Notice that MNLI is totally out-of-domain regarding the task of QA

Improving the prediction confidence

Improving the prediction confidence

- Selective QA: If our model can choose to answer only the k percentage of examples it is most confident about (the coverage), what F1 can it achieve?
  - Examples are ranked by the confidence score of a model
  - QA -> posterior probability of the answer span
  - NLI -> posterior probability associated with “Entailment”

Improving the prediction confidence

- Selective QA: If our model can choose to answer only the k percentage of examples it is most confident about (the coverage), what F1 can it achieve?
  - Examples are ranked by the confidence score of a model
  - QA -> posterior probability of the answer span
  - NLI -> posterior probability associated with “Entailment”

- Experimental setup:

Improving the prediction confidence

- Selective QA: If our model can choose to answer only the k percentage of examples it is most confident about (the coverage), what F1 can it achieve?
  - Examples are ranked by the confidence score of a model
  - QA -> posterior probability of the answer span
  - NLI -> posterior probability associated with “Entailment”

- Experimental setup:
  - Train a QA model on Natural Questions (Kwiatkowski et al. 2019) and test it on NQ and 4 out-of-domain datasets.

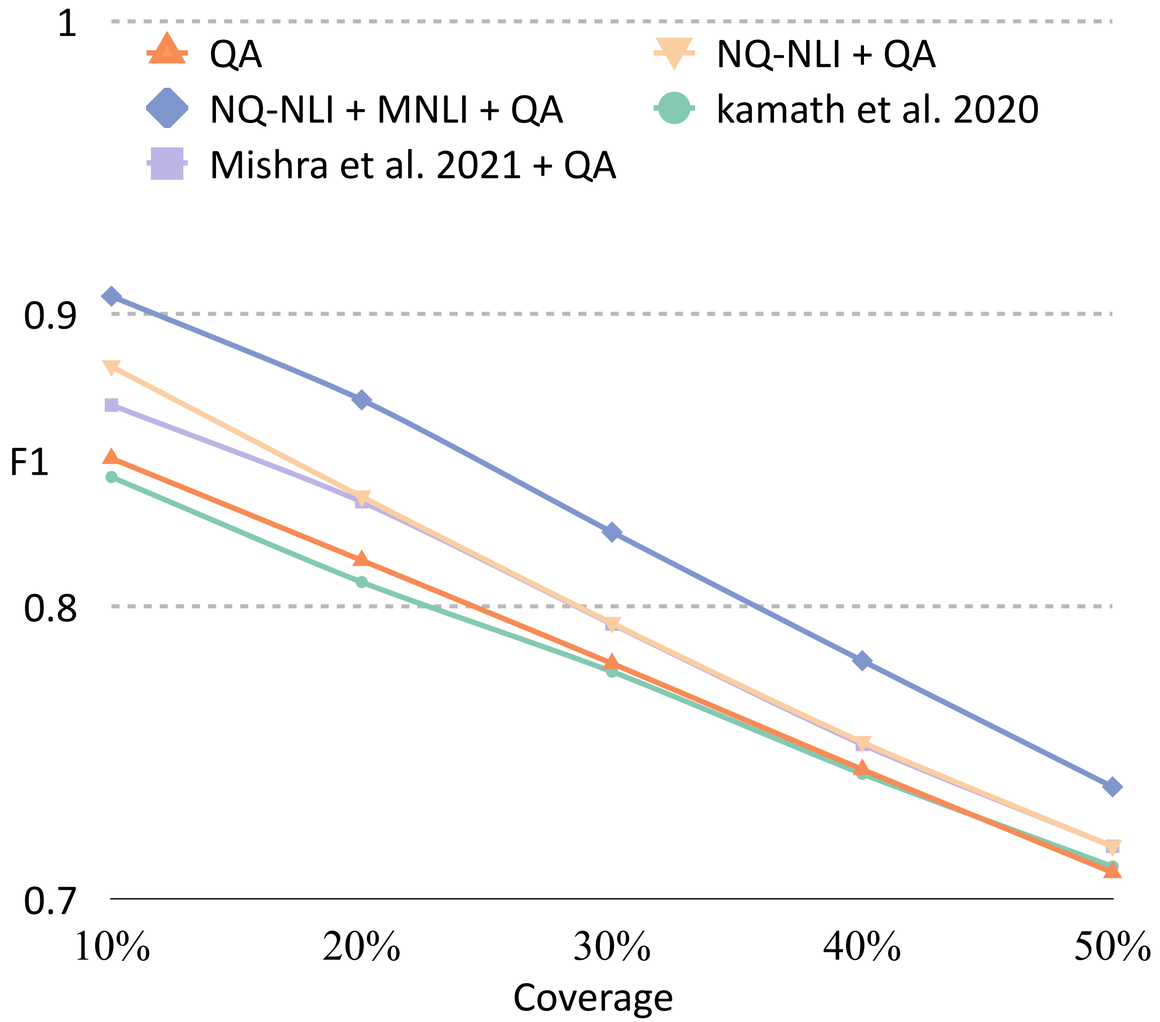
Improving the prediction confidence

- Selective QA: If our model can choose to answer only the k percentage of examples it is most confident about (the coverage), what F1 can it achieve?
  - Examples are ranked by the confidence score of a model
  - QA -> posterior probability of the answer span
  - NLI -> posterior probability associated with “Entailment”

- Experimental setup:
  - Train a QA model on Natural Questions (Kwiatkowski et al. 2019) and test it on NQ and 4 out-of-domain datasets.
  - Train a NLI model using the generated NLI pairs from Natural Questions and use it to verify the predictions from the previous step

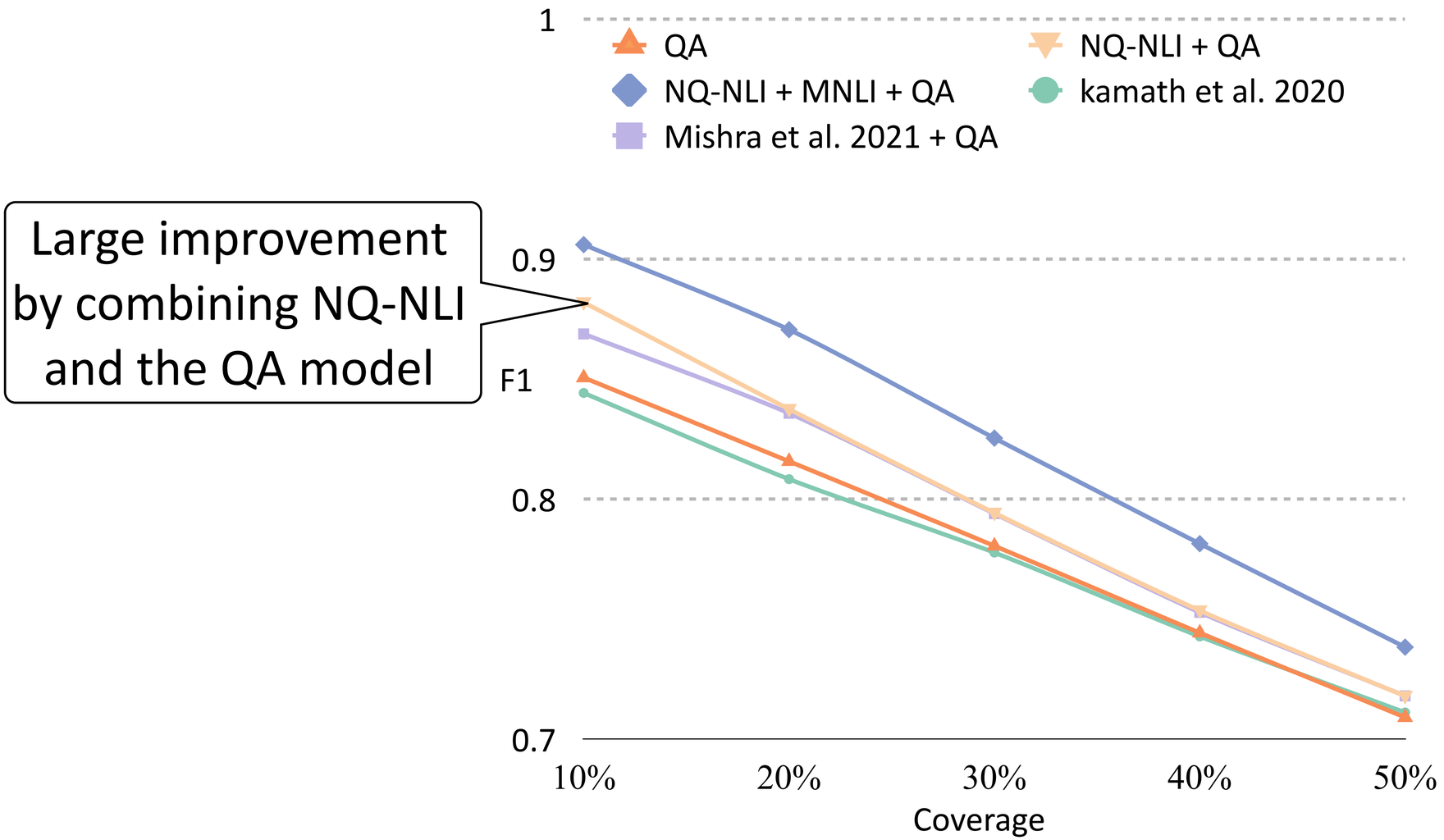
Improving the prediction confidence

- Results:



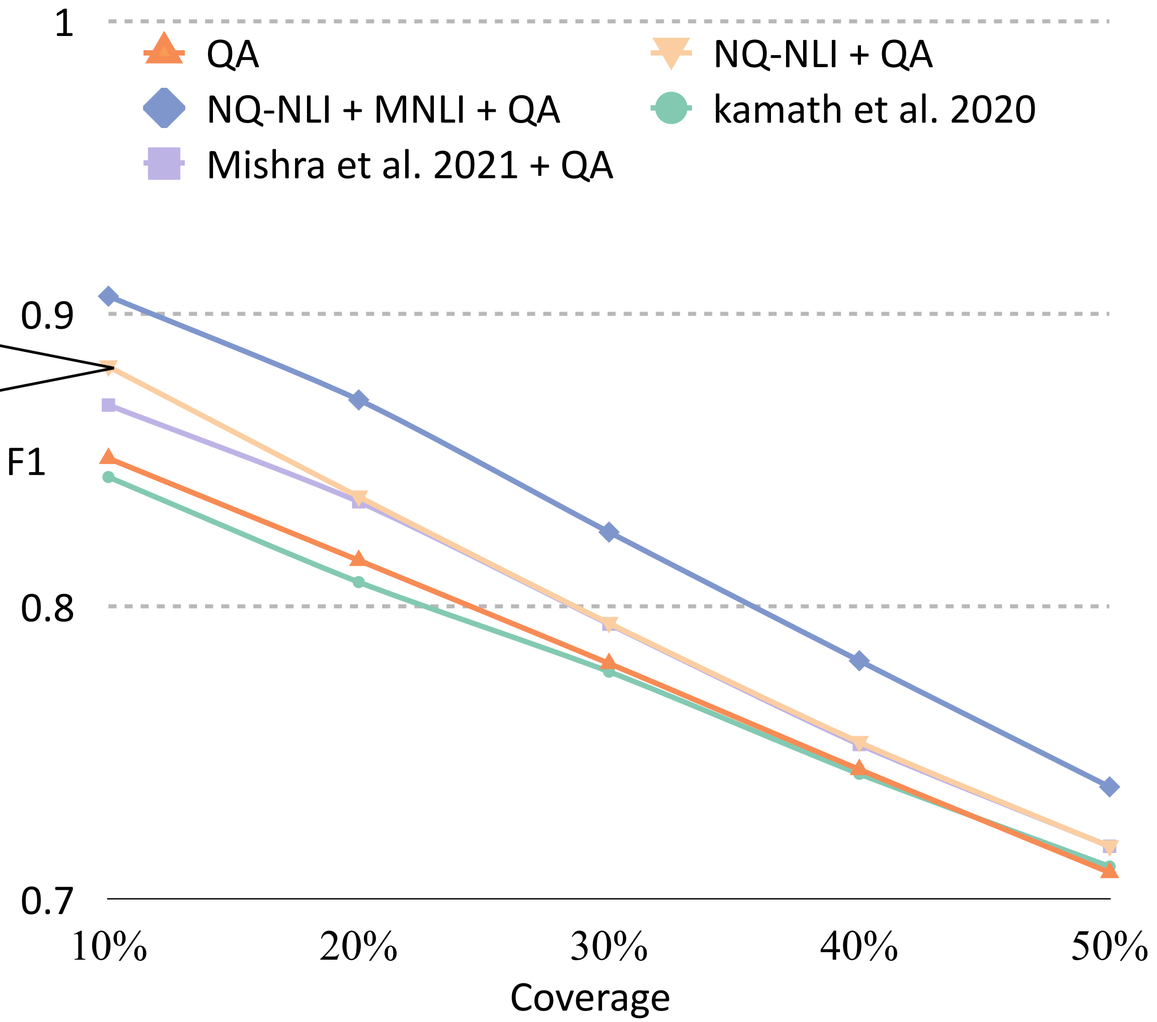
Improving the prediction confidence

- Results:



Improving the prediction confidence

- Results:

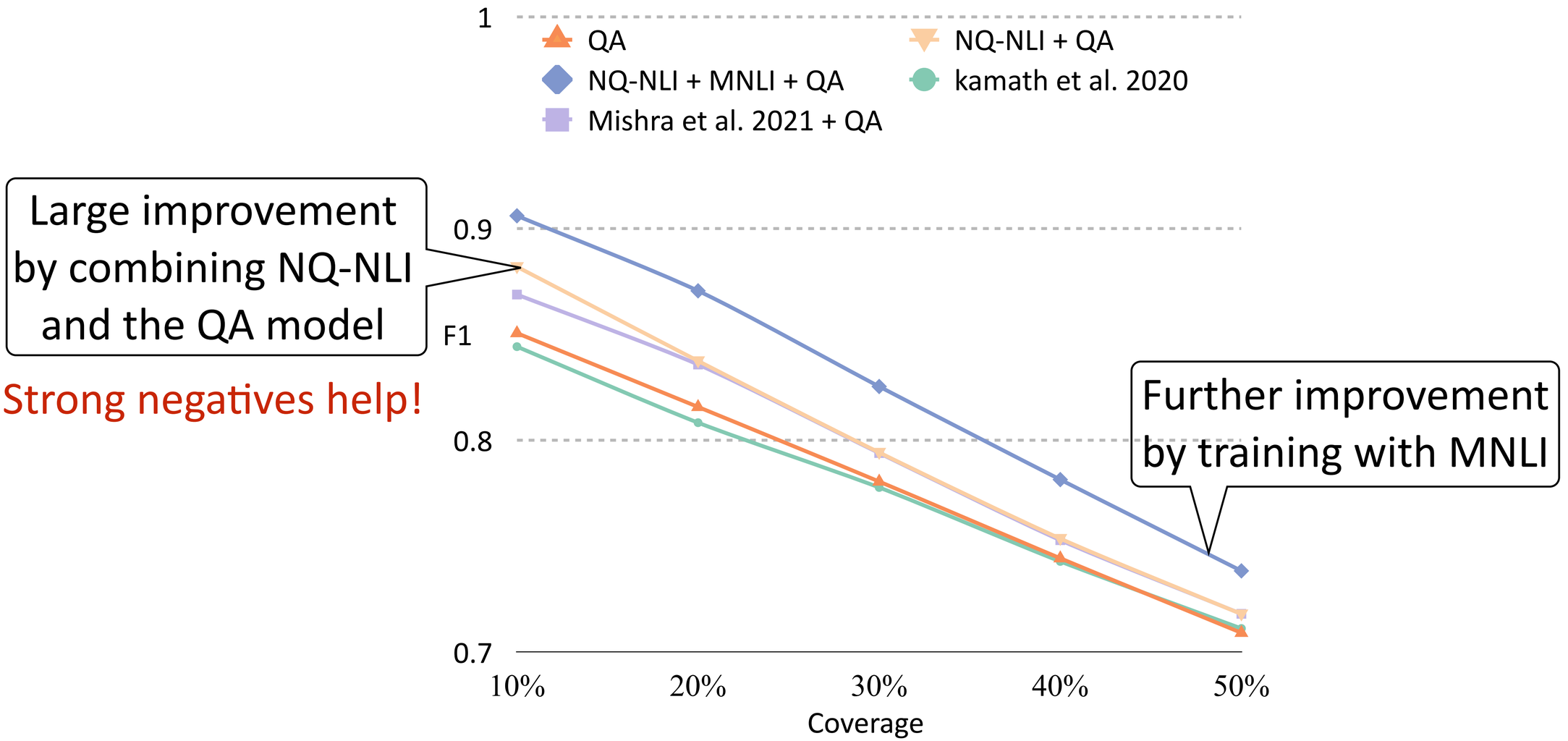


Large improvement by combining NQ-NLI and the QA model

Strong negatives help!

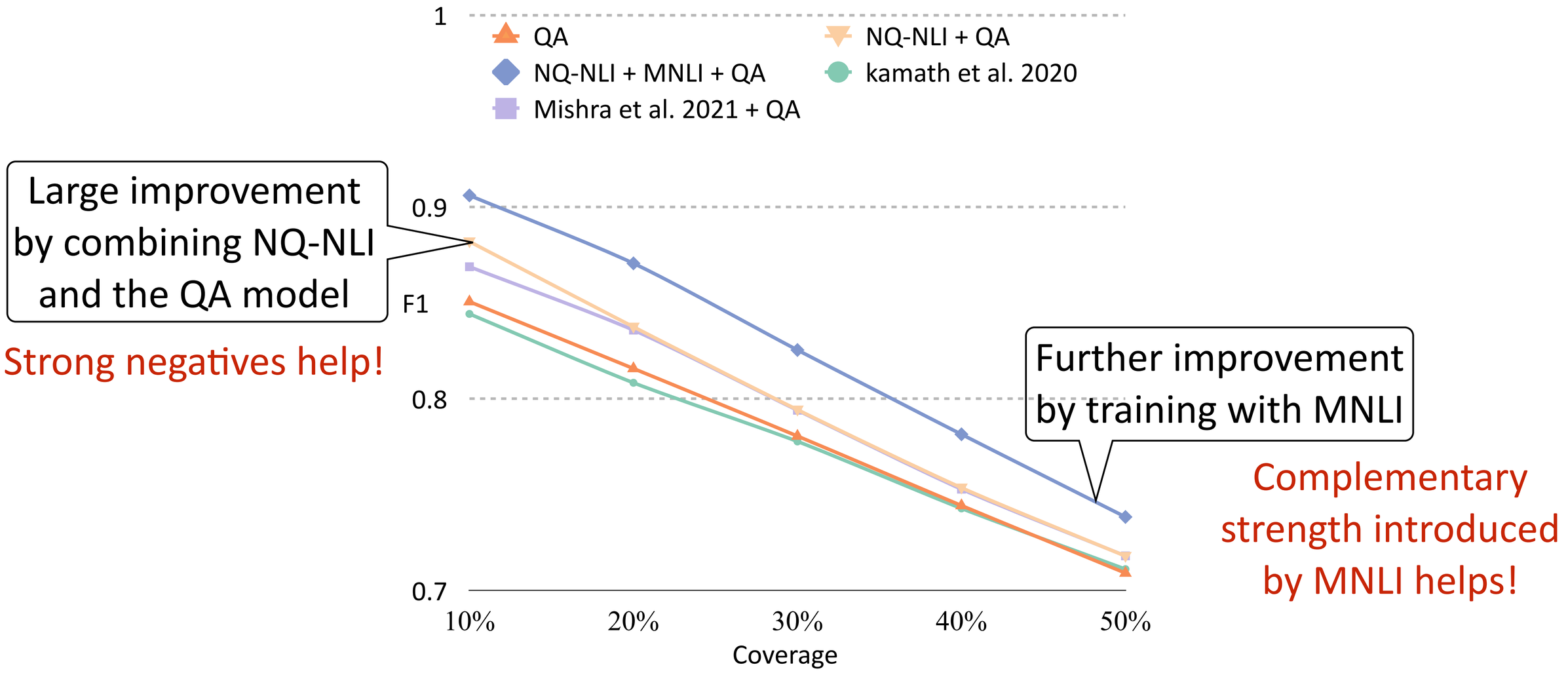
Improving the prediction confidence

- Results:



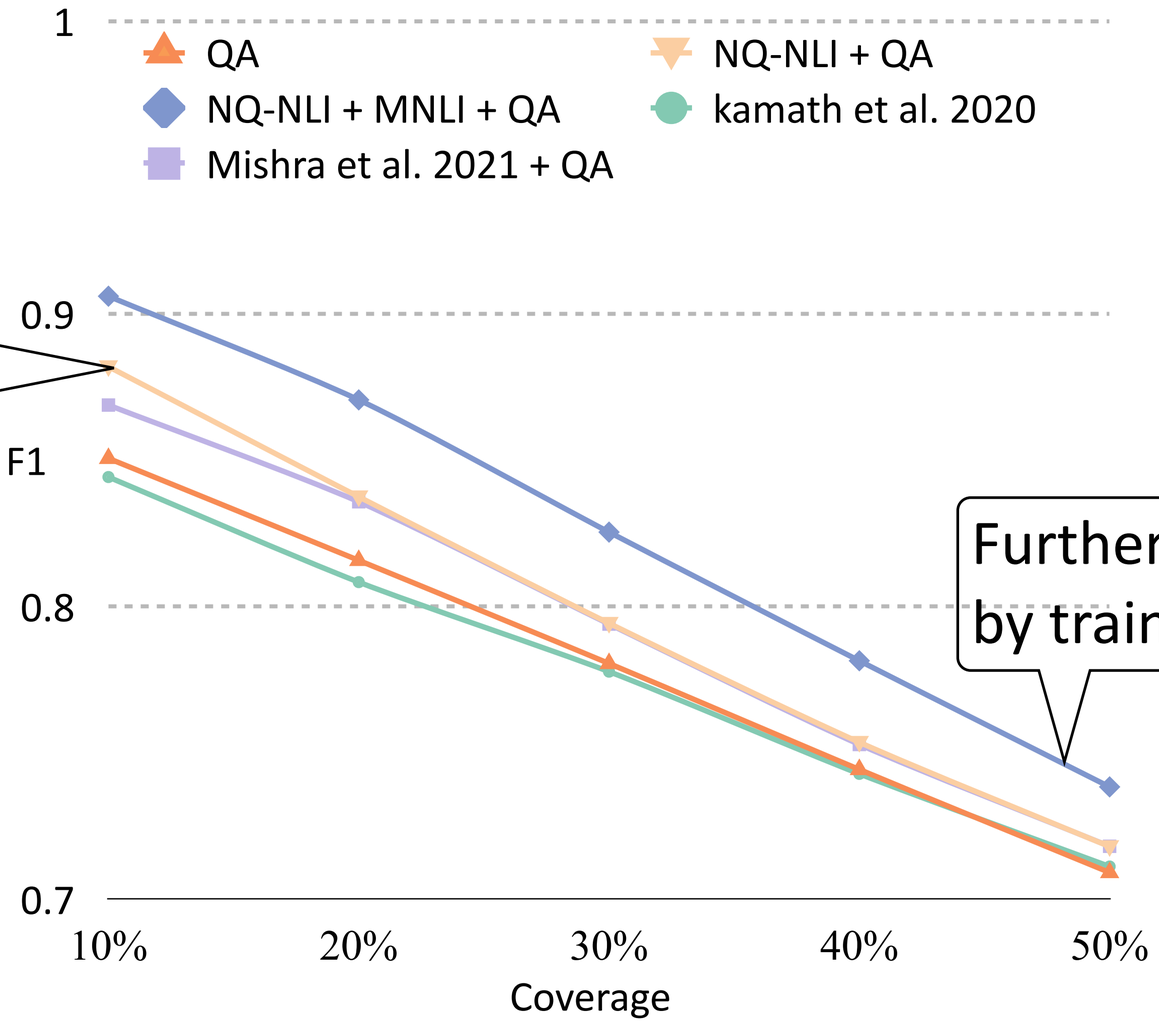
Improving the prediction confidence

- Results:



Improving the prediction confidence

- Results:



Large improvement by combining NQ-NLI and the QA model

Strong negatives help!

For more results and ablations, check out our paper!

Further improvement by training with MNLI

Complementary strength introduced by MNLI helps!

Disagreement between NLI and QA

- Right answer for the wrong reason (25% in NQ):

Disagreement between NLI and QA

- Right answer for the wrong reason (25% in NQ):

Question: When was Clash Royale released in the US?

Answer: The game Clash Royale was released globally on **March 2 , 2016**

Disagreement between NLI and QA

- Right answer for the wrong reason (25% in NQ):

Question: When was Clash Royale released in the US?

Answer: The game Clash Royale was released globally on **March 2 , 2016**

Globally —> US? Need further evidence

Disagreement between NLI and QA

- Right answer for the wrong reason (25% in NQ):

Question: When was Clash Royale released in the US?

Answer: The game Clash Royale was released globally on **March 2 , 2016**

Globally —> US? Need further evidence

Question: Who plays the bad guy in the Good Place?

Answer: The series The Good Place focuses on Eleanor Shellstrop (Kristen Bell) , a woman who wakes up in the afterlife and is introduced by Michael (**Ted Danson**) to The Good Place '' …

Disagreement between NLI and QA

- Right answer for the wrong reason (25% in NQ):

Question: When was Clash Royale released in the US?

Answer: The game Clash Royale was released globally on **March 2 , 2016**

Globally —> US? Need further evidence

Question: Who plays the bad guy in the Good Place?

Answer: The series The Good Place focuses on Eleanor Shellstrop (Kristen Bell) , a woman who wakes up in the afterlife and is introduced by Michael (**Ted Danson**) to The Good Place '' …

Is Michael the bad guy? Need to check

Takeaways

Takeaways

- Existing QA datasets encourage models to return answers when the context does not actually contain sufficient information
  - Fully verifying the answer is a challenging direction

Takeaways

- Existing QA datasets encourage models to return answers when the context does not actually contain sufficient information
  - Fully verifying the answer is a challenging direction

- The proposed approach is helpful to locate the information mismatches between the question and the supporting context

Takeaways

- Existing QA datasets encourage models to return answers when the context does not actually contain sufficient information
  - Fully verifying the answer is a challenging direction

- The proposed approach is helpful to locate the information mismatches between the question and the supporting context

Thank you!

Q&A