

Is Multihop Reasoning in DiRe Condition? Measuring and Reducing Disconnected Reasoning



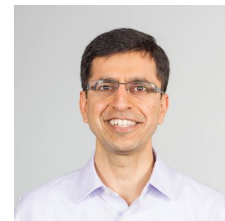
Harsh Trivedi



Niranjan Balasubramanian



Tushar Khot



Ashish Sabharwal



Stony Brook University



**ALLEN INSTITUTE
for ARTIFICIAL INTELLIGENCE**

Multihop reasoning datasets have artifacts

Compositional Questions Do Not Necessitate Multi-hop Reasoning

**Sewon Min^{*1}, Eric Wallace^{*2}, Sameer Singh³,
Matt Gardner², Hannaneh Hajishirzi^{1,2}, Luke Zettlemoyer¹**

¹University of Washington

²Allen Institute for Artificial Intelligence

³University of California, Irvine

sewon@cs.washington.edu, ericw@allenai.org

Understanding Dataset Design Choices for Multi-hop Reasoning

Jifan Chen and Greg Durrett
The University of Texas at Austin
{jfchen, gdurrett}@cs.utexas.edu

Avoiding Reasoning Shortcuts: Adversarial Evaluation, Training, and Model Development for Multi-Hop QA

Yichen Jiang and Mohit Bansal
UNC Chapel Hill
{yichenj, mbansal}@cs.unc.edu

Multihop reasoning datasets have artifacts

Compositional Questions Do Not Necessitate Multi-hop Reasoning

Sewon Min^{*1}, Eric Wallace^{*2}, Sameer Singh³,
Matt Gardner², Hannaneh Hajishirzi^{1,2}, Luke Zettlemoyer¹

¹University of Washington

²Allen Institute for Artificial Intelligence

³University of California, Irvine

sewon@cs.washington.edu, ericw@allenai.org

Undesirable reasoning models, ones that **by design** *cannot* do multihop reasoning, *can* achieve high scores because of dataset biases.

Understanding Dataset Design Choices for Multi-hop Reasoning

Jifan Chen and Greg Durrett
The University of Texas at Austin
{jfchen, gdurrett}@cs.utexas.edu

Avoiding Reasoning Shortcuts: Adversarial Evaluation, Training, and Model Development for Multi-Hop QA

Yichen Jiang and Mohit Bansal
UNC Chapel Hill
{yichenj, mbansal}@cs.unc.edu

Multihop reasoning datasets have artifacts

Compositional Questions Do Not Necessitate Multi-hop Reasoning

Sewon Min^{*1}, Eric Wallace^{*2}, Sameer Singh³,
Matt Gardner², Hannaneh Hajishirzi^{1,2}, Luke Zettlemoyer¹

¹University of Washington

²Allen Institute for Artificial Intelligence

³University of California, Irvine

sewon@cs.washington.edu, ericw@allenai.org

Understanding Dataset Design Choices for Multi-hop Reasoning

Jifan Chen and Greg Durrett
The University of Texas at Austin
{jfchen, gdurrett}@cs.utexas.edu

Avoiding Reasoning Shortcuts: Adversarial Evaluation, Training, and Model Development for Multi-Hop QA

Yichen Jiang and Mohit Bansal
UNC Chapel Hill
{yichenj, mbansal}@cs.unc.edu

Undesirable reasoning models, ones that **by design cannot** do multihop reasoning, *can* achieve high scores because of dataset biases.



The *existence of models* that *can* cheat.

Open questions

Compositional Questions Do Not Necessitate Multi-hop Reasoning

Sewon Min^{*1}, Eric Wallace^{*2}, Sameer Singh³,
Matt Gardner², Hannaneh Hajishirzi^{1,2}, Luke Zettlemoyer¹

¹University of Washington

²Allen Institute for Artificial Intelligence

³University of California, Irvine

sewon@cs.washington.edu, ericw@allenai.org

Understanding Dataset Design Choices for Multi-hop Reasoning

Jifan Chen and Greg Durrett
The University of Texas at Austin
{jfchen, gdurrett}@cs.utexas.edu

Avoiding Reasoning Shortcuts: Adversarial Evaluation, Training, and Model Development for Multi-Hop QA

Yichen Jiang and Mohit Bansal
UNC Chapel Hill
{yichenj, mbansal}@cs.unc.edu

Undesirable reasoning models, ones that **by design** *cannot* do multihop reasoning, *can* achieve high scores because of dataset biases.



The *existence of models* that *can* cheat.



To what extent a *given model* cheats?

Open questions

Compositional Questions Do Not Necessitate Multi-hop Reasoning

Sewon Min^{*1}, Eric Wallace^{*2}, Sameer Singh³,
Matt Gardner², Hannaneh Hajishirzi^{1,2}, Luke Zettlemoyer¹

¹University of Washington

²Allen Institute for Artificial Intelligence

³University of California, Irvine

sewon@cs.washington.edu, ericw@allenai.org

Understanding Dataset Design Choices for Multi-hop Reasoning

Jifan Chen and Greg Durrett
The University of Texas at Austin
{jfchen, gdurrett}@cs.utexas.edu

Avoiding Reasoning Shortcuts: Adversarial Evaluation, Training, and Model Development for Multi-Hop QA

Yichen Jiang and Mohit Bansal
UNC Chapel Hill
{yichenj, mbansal}@cs.unc.edu

Undesirable reasoning models, ones that **by design** *cannot* do multihop reasoning, *can* achieve high scores because of dataset biases.



The *existence of models* that *can* cheat.



To what extent a *given model* cheats?



Has there been real progress in the multihop reasoning skill of models?

Open questions

Compositional Questions Do Not Necessitate Multi-hop Reasoning

Sewon Min^{*1}, Eric Wallace^{*2}, Sameer Singh³,
Matt Gardner², Hannaneh Hajishirzi^{1,2}, Luke Zettlemoyer¹

¹University of Washington

²Allen Institute for Artificial Intelligence

³University of California, Irvine

sewon@cs.washington.edu, ericw@allenai.org

Understanding Dataset Design Choices for Multi-hop Reasoning

Jifan Chen and Greg Durrett
The University of Texas at Austin
{jfchen, gdurrett}@cs.utexas.edu

Avoiding Reasoning Shortcuts: Adversarial Evaluation, Training, and Model Development for Multi-Hop QA

Yichen Jiang and Mohit Bansal
UNC Chapel Hill
{yichenj, mbansal}@cs.unc.edu

Undesirable reasoning models, ones that **by design** *cannot* do multihop reasoning, *can* achieve high scores because of dataset biases.



The *existence of models* that *can* cheat.



To what extent a *given model* cheats?



Has there been real progress in the multihop reasoning skill of models?



What can we do to reduce cheatability of these datasets?

Outline of this work

We address these for reading comprehension multihop QA with annotated supporting facts.

Outline of this work

We address these for reading comprehension multihop QA with annotated supporting facts.



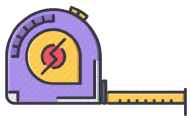
- **Introduce Disconnected Reasoning (DiRe)**
 - a model-agnostic characterization of a form of non-multihop reasoning.

Outline of this work

We address these for reading comprehension multihop QA with annotated supporting facts.



- **Introduce Disconnected Reasoning (DiRe)**
 - a model-agnostic characterization of a form of non-multihop reasoning.



- **Measure Disconnected Reasoning**
 - Done by the given model.
 - Possible on the dataset.

Outline of this work

We address these for reading comprehension multihop QA with annotated supporting facts.



- **Introduce Disconnected Reasoning (DiRe)**
 - a model-agnostic characterization of a form of non-multihop reasoning.



- **Measure Disconnected Reasoning**
 - Done by the given model.
 - Possible on the dataset.



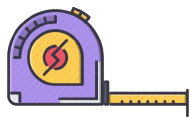
- **Reduce Disconnected Reasoning**
 - Automatic transformation to reduce dataset cheatability.

Outline of this work

We address these for reading comprehension multihop QA with annotated supporting facts.



- **Introduce Disconnected Reasoning (DiRe)**
 - a model-agnostic characterization of a form of non-multihop reasoning.



- **Measure** Disconnected Reasoning
 - Done by the given model.
 - Possible on the dataset.



- **Reduce** Disconnected Reasoning
 - Automatic transformation to reduce dataset cheatability.

Example Multifact Question



Which country got independence when the cold war started?



The cold war started in 1947.



India got independence from UK in 1947.

Example Multifact Question



Which country got independence when the cold war started?



The war started in 1950.



The cold war started in 1947.



France finally got its independence.



India got independence from UK in 1947.



30 countries were involved in World War 2.

Example Multifact Question



Which country got independence when the cold war started?



The war started in 1950.



The cold war started in 1947.



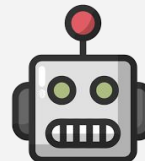
France finally got its independence.



India got independence from UK in 1947.



30 countries were involved in World War 2.



Example Multifact Question



Which country got independence when the cold war started?



The war started in 1950.



The cold war started in 1947.



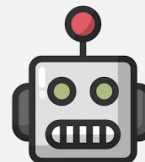
France finally got its independence.



India got independence from UK in 1947.



30 countries were involved in World War 2.



Answer Test
India

Example Multifact Question



Which country got independence when the cold war started?



The war started in 1950.



The cold war started in 1947.



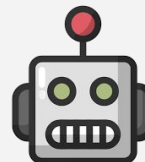
France finally got its independence.



India got independence from UK in 1947.



30 countries were involved in World War 2.



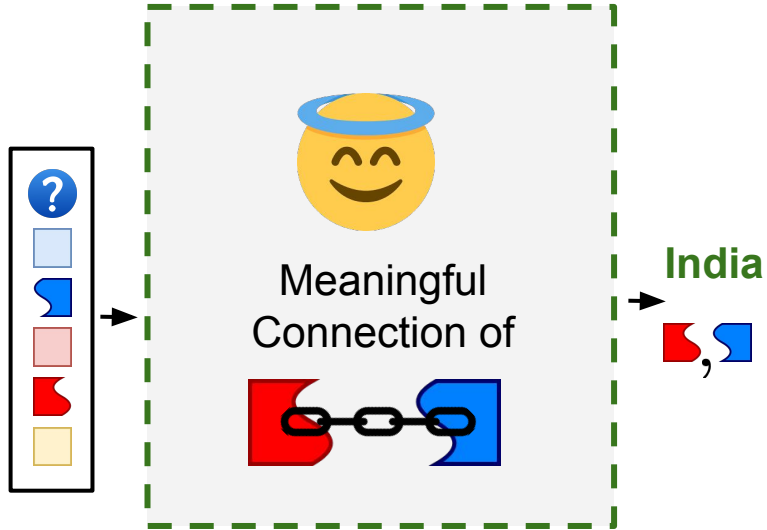
Answer Test

India

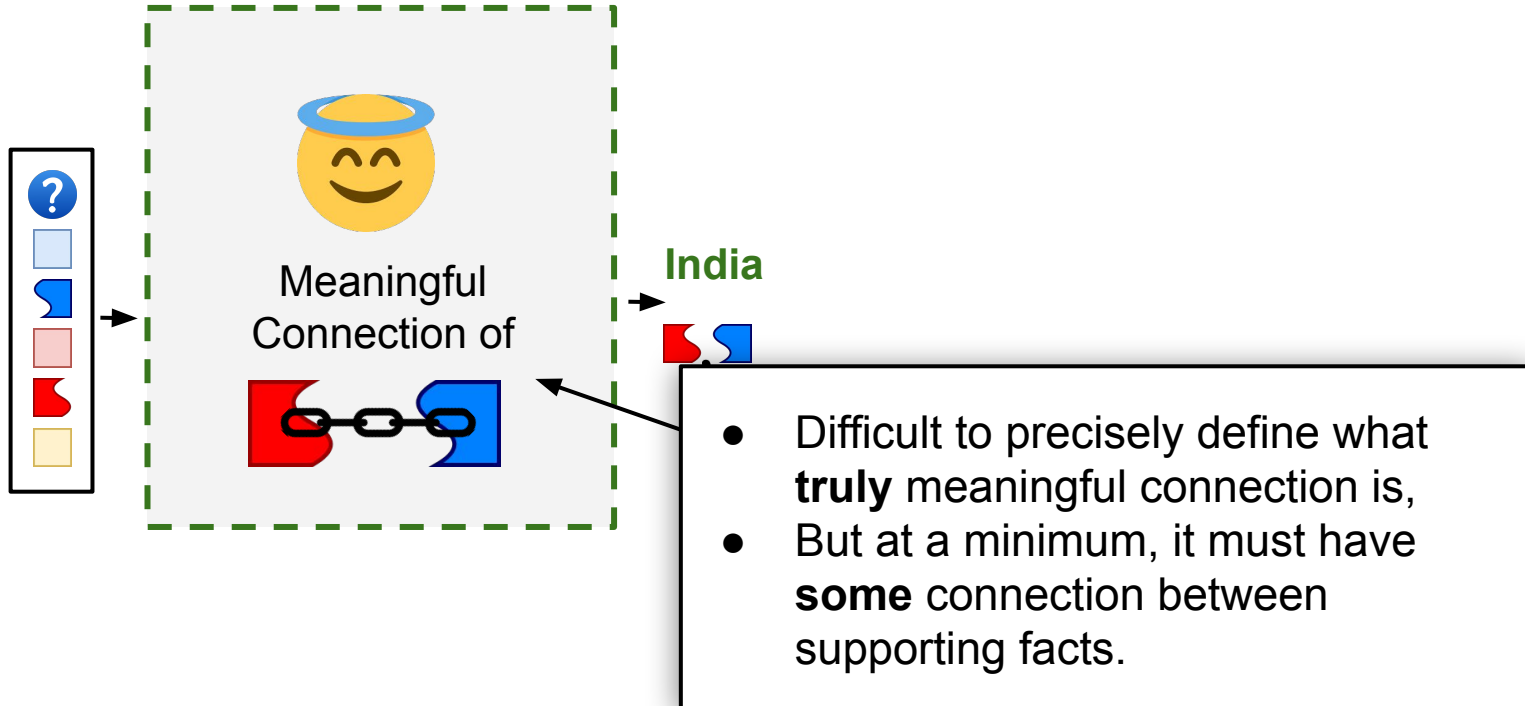
**Supporting
Fact (SF) Test**



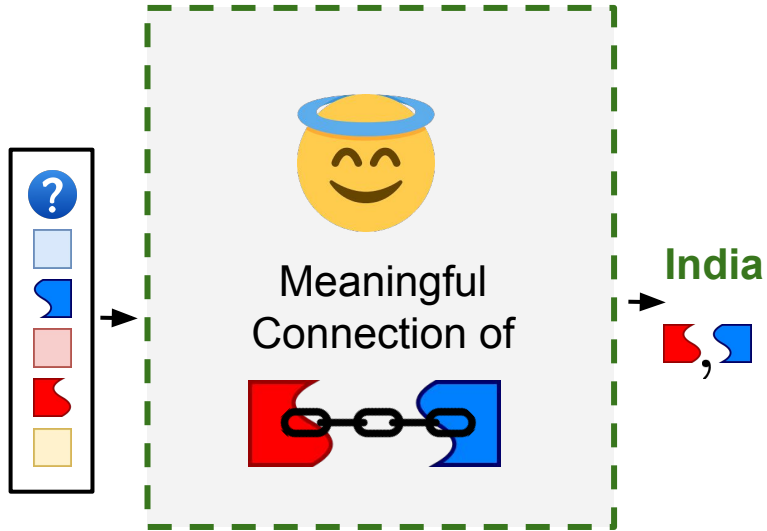
Multifact Reasoning



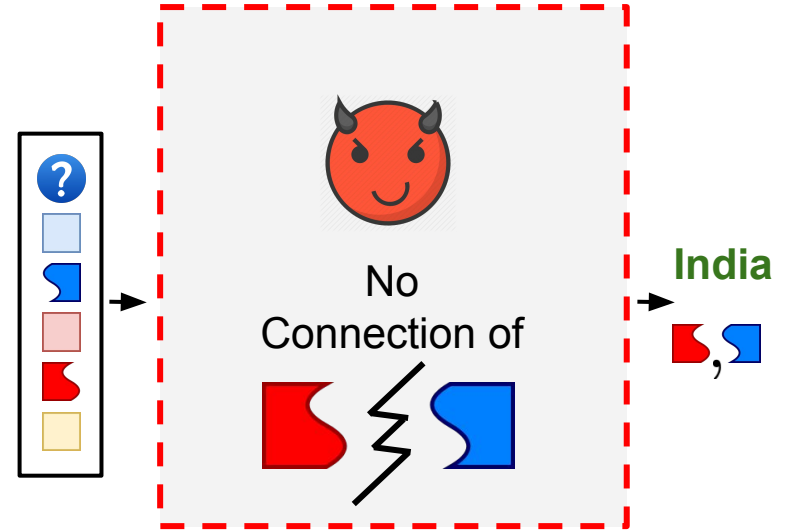
Multifact Reasoning



Multifact Reasoning

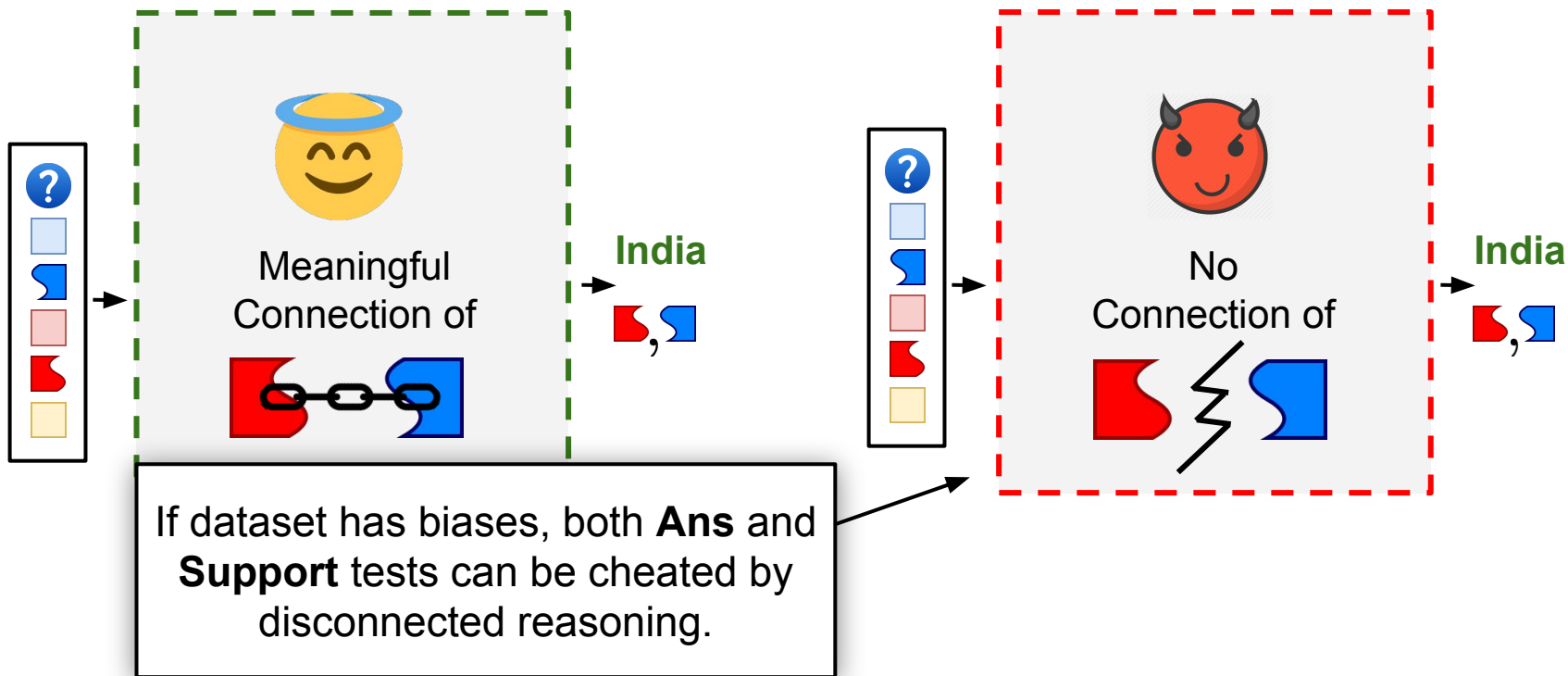


Disconnected Reasoning



Multifact Reasoning

Disconnected Reasoning



Cheating by Disconnected Reasoning



Which country got independence when the cold war started?



The war started in 1950.



The cold war started in 1947.



France finally got its independence.



India got independence from UK in 1947.



30 countries were involved in World War 2.



Answer Test

India

**Supporting
Fact (SF) Test**



Cheating by Disconnected Reasoning



Which country got independence when the **cold war started**?



The war started in 1950.



The **cold war started** in 1947.



France finally got its independence.



India got independence from UK in 1947.



30 countries were involved in World War 2.



Answer

SF



Answer Test

India

Supporting
Fact (SF) Test



Cheating by Disconnected Reasoning



Which **country got independence** when the cold war started?



The war started in 1950.



~~The cold war started in 1947.~~



France finally got its independence.



India got independence from UK in **1947**.



30 countries were involved in World War 2.



Answer SF

[India]



Answer Test

India

Supporting
Fact (SF) Test



Cheating by Disconnected Reasoning



Which country got independence when the cold war started?



The war started in 1950.



The cold war started in 1947.



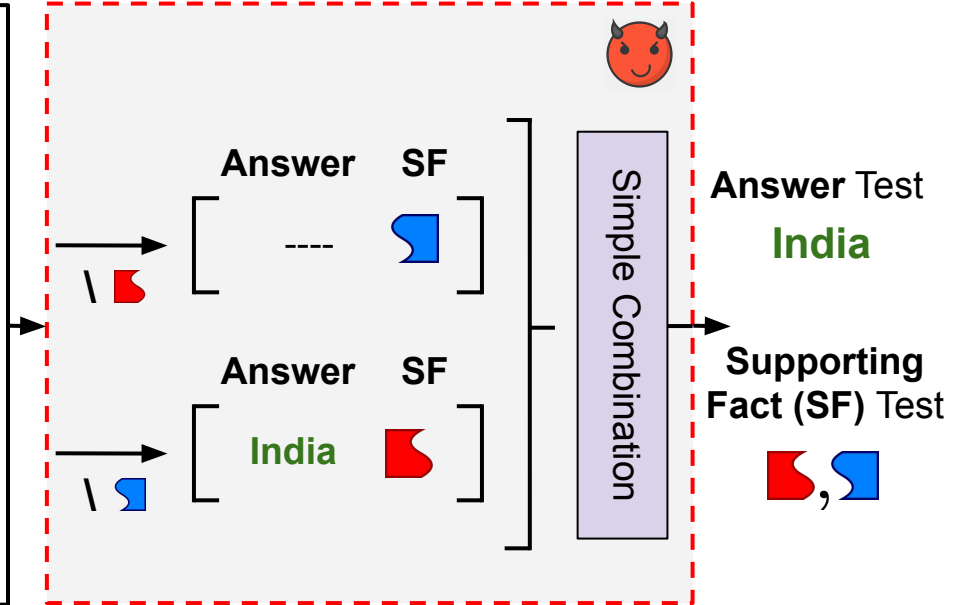
France finally got its independence.



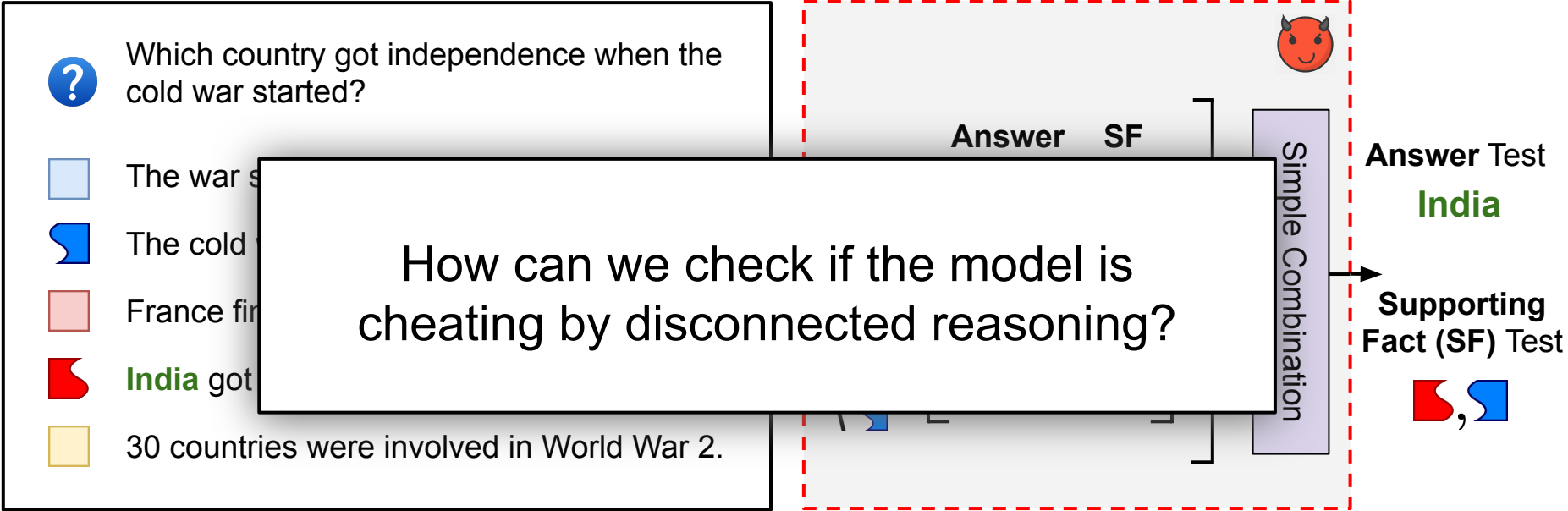
India got independence from UK in 1947.



30 countries were involved in World War 2.



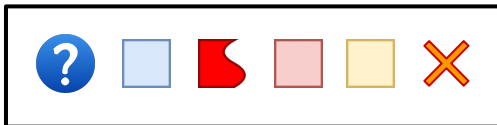
Cheating by Disconnected Reasoning



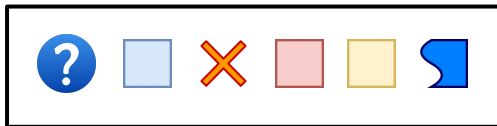
Disconnected Reasoning (DiRe) Condition

Disconnected Reasoning (DiRe) Condition

Doesn't contain



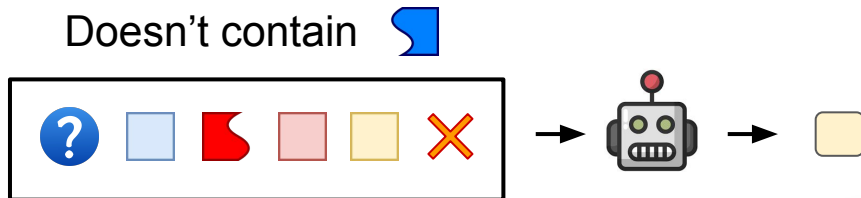
Consider different parts of the input such that no part contains all the supporting facts.



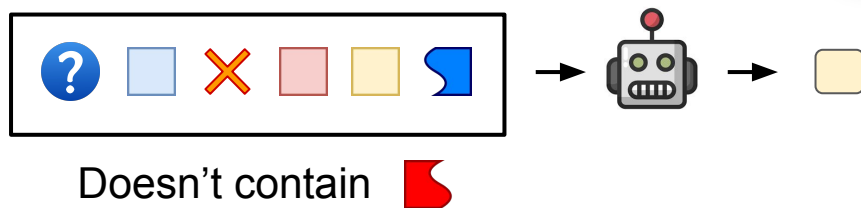
Doesn't contain



Disconnected Reasoning (DiRe) Condition






Suppose we run the model on each part independently.



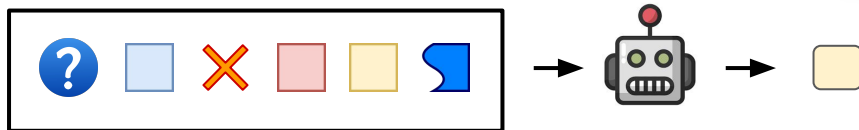
Disconnected Reasoning (DiRe) Condition

Doesn't contain 



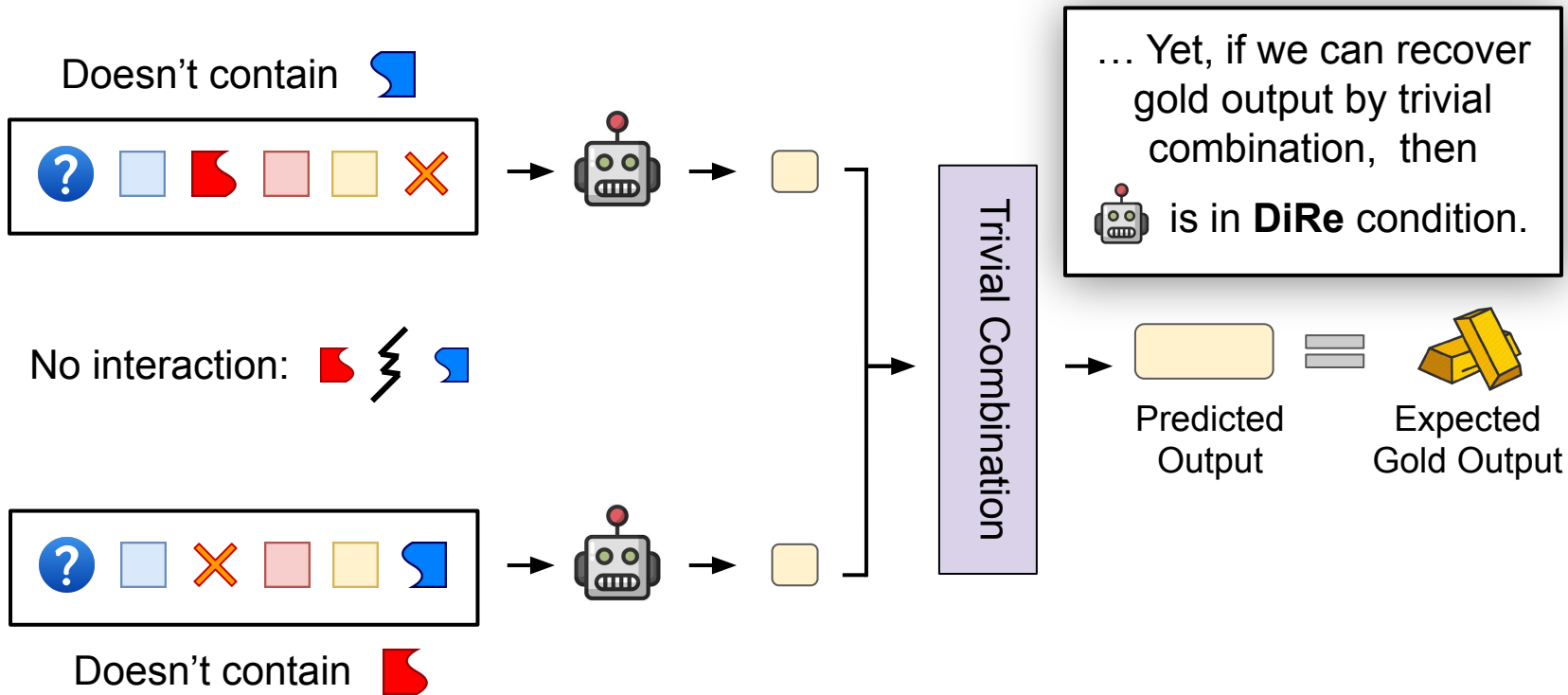
No interaction:   

The model, by design, cannot connect information from  and  together.

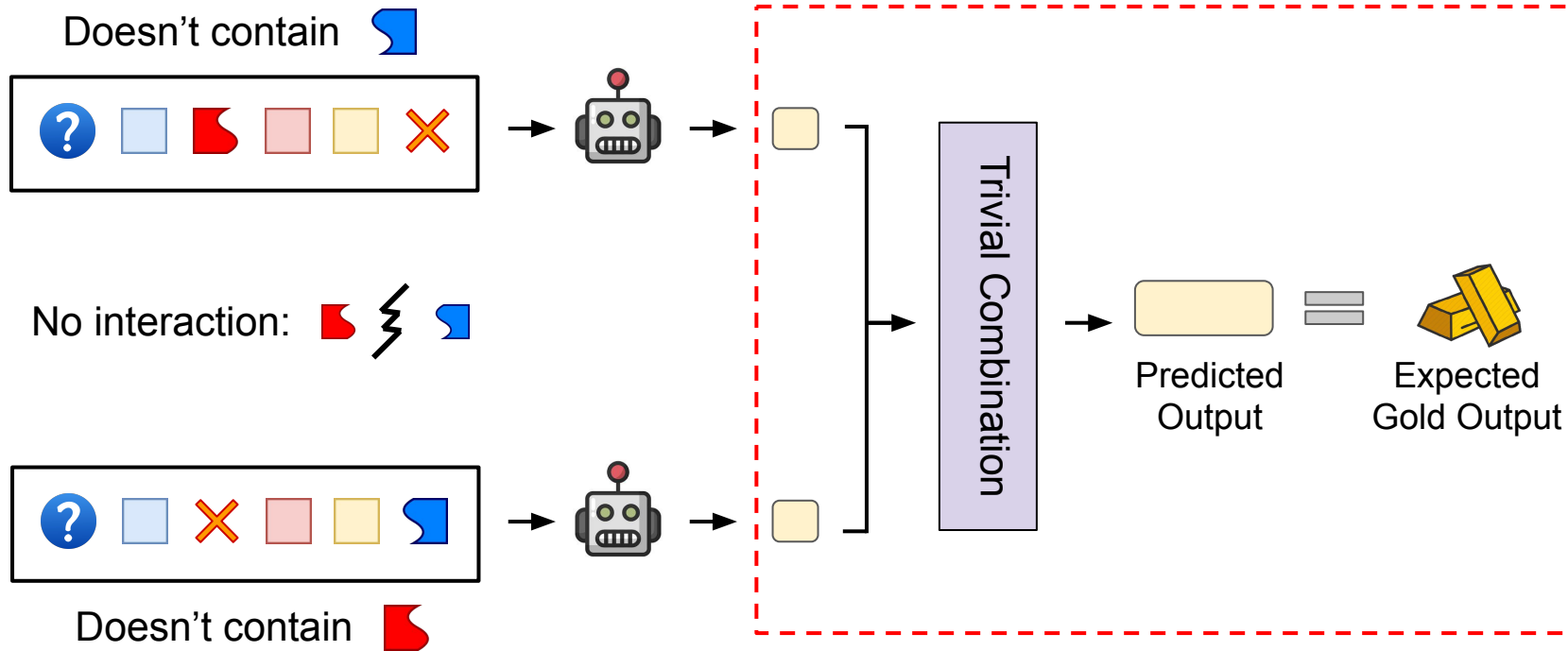


Doesn't contain 

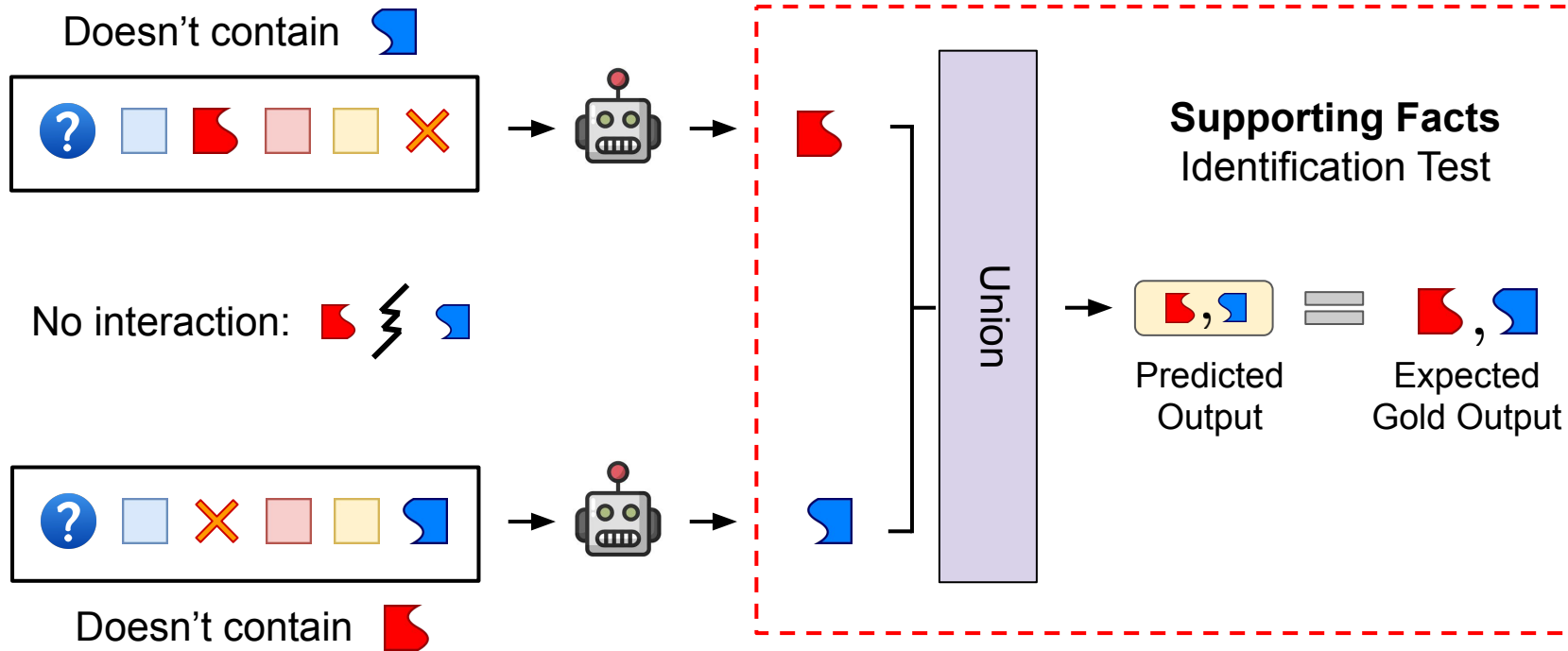
Disconnected Reasoning (DiRe) Condition



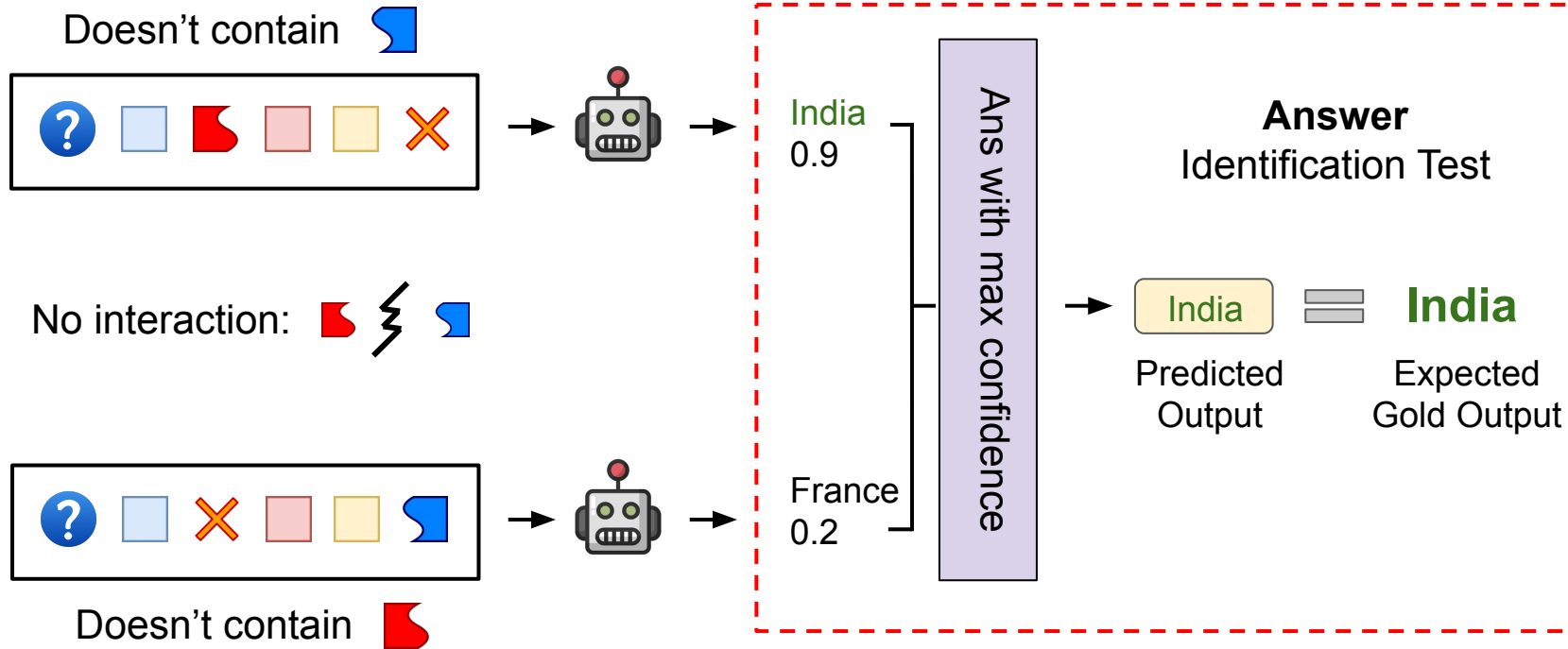
Disconnected Reasoning (DiRe) Condition



Disconnected Reasoning (DiRe) Condition



Disconnected Reasoning (DiRe) Condition

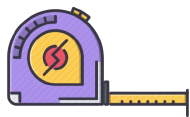


Outline of this work

We address these for reading comprehension multihop QA with annotated supporting facts.



- **Introduce Disconnected Reasoning (DiRe)**
 - a model-agnostic characterization of a form of non-multihop reasoning.

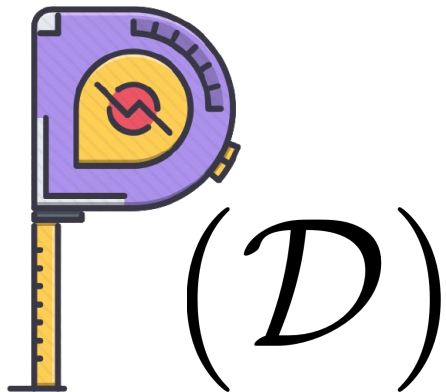


- **Measure Disconnected Reasoning**
 - Done by the given model.
 - Possible on the dataset.



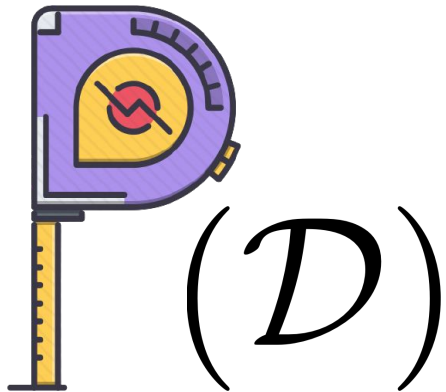
- **Reduce Disconnected Reasoning**
 - Automatic transformation to reduce dataset cheatability.

Measuring Disconnected Reasoning



DiRe Probing dataset of \mathcal{D}

Measuring Disconnected Reasoning

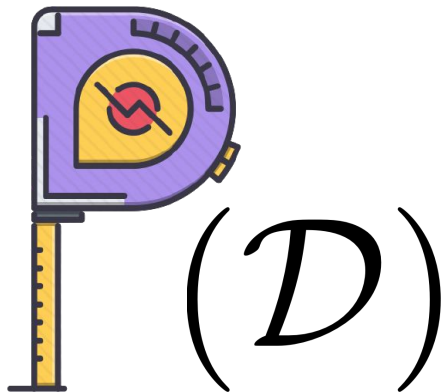


DiRe **P**robing dataset of \mathcal{D}

Measure disconnected reasoning

- of model 🤖 on dataset \mathcal{D} .
- possible on the dataset \mathcal{D} .

Measuring Disconnected Reasoning



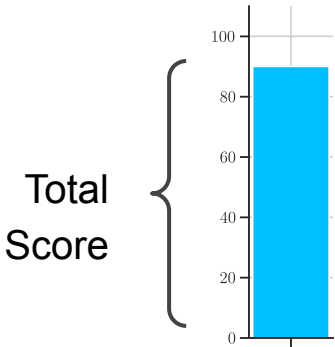
DiRe **P**robing dataset of \mathcal{D}

Measure disconnected reasoning

- of model 🤖 on dataset \mathcal{D} .
- possible on the dataset \mathcal{D} .

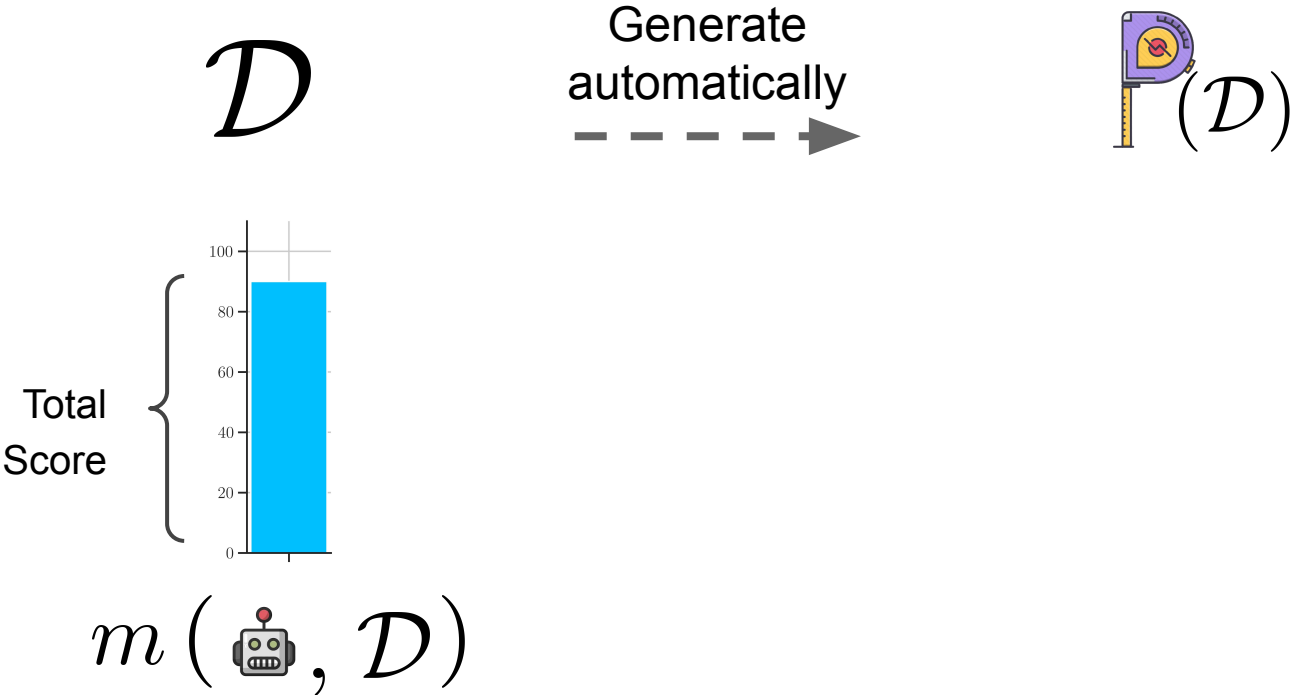
Measure Disconnected Reasoning of Model

\mathcal{D}

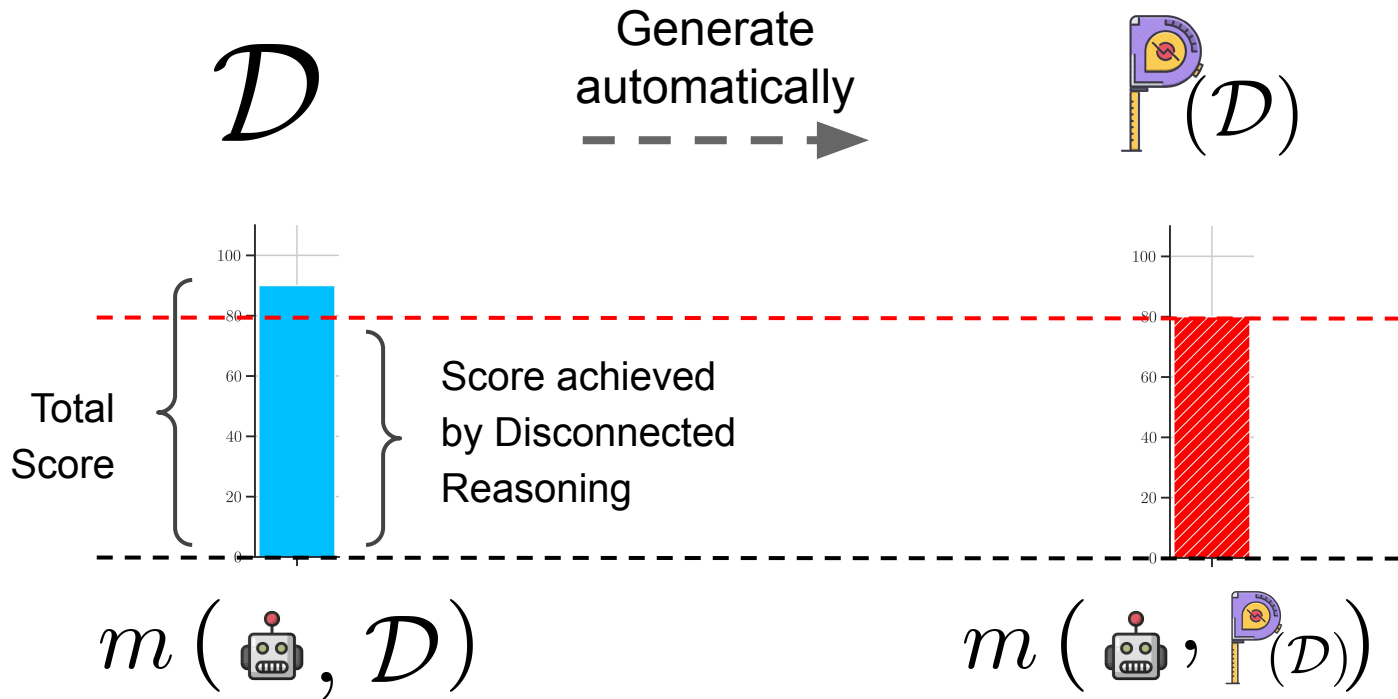


$m(\text{robot icon}, \mathcal{D})$

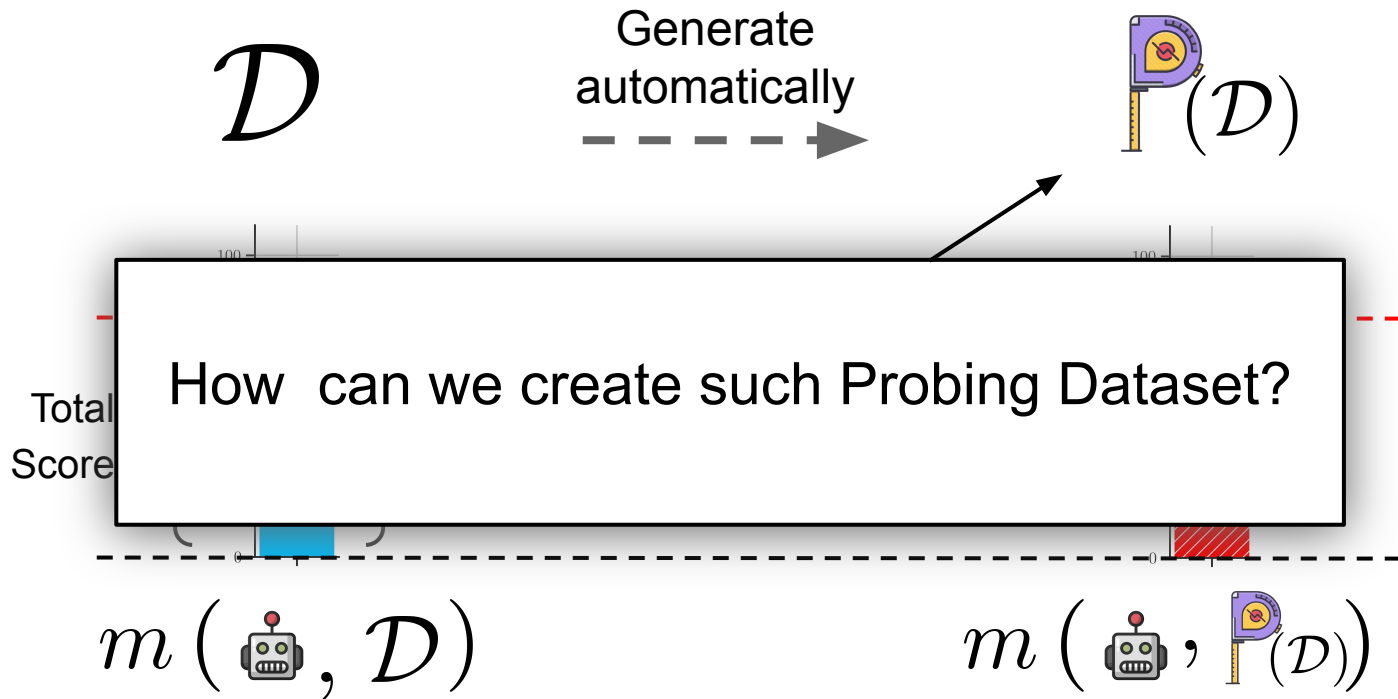
Measure Disconnected Reasoning of Model



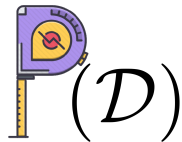
Measure Disconnected Reasoning of Model



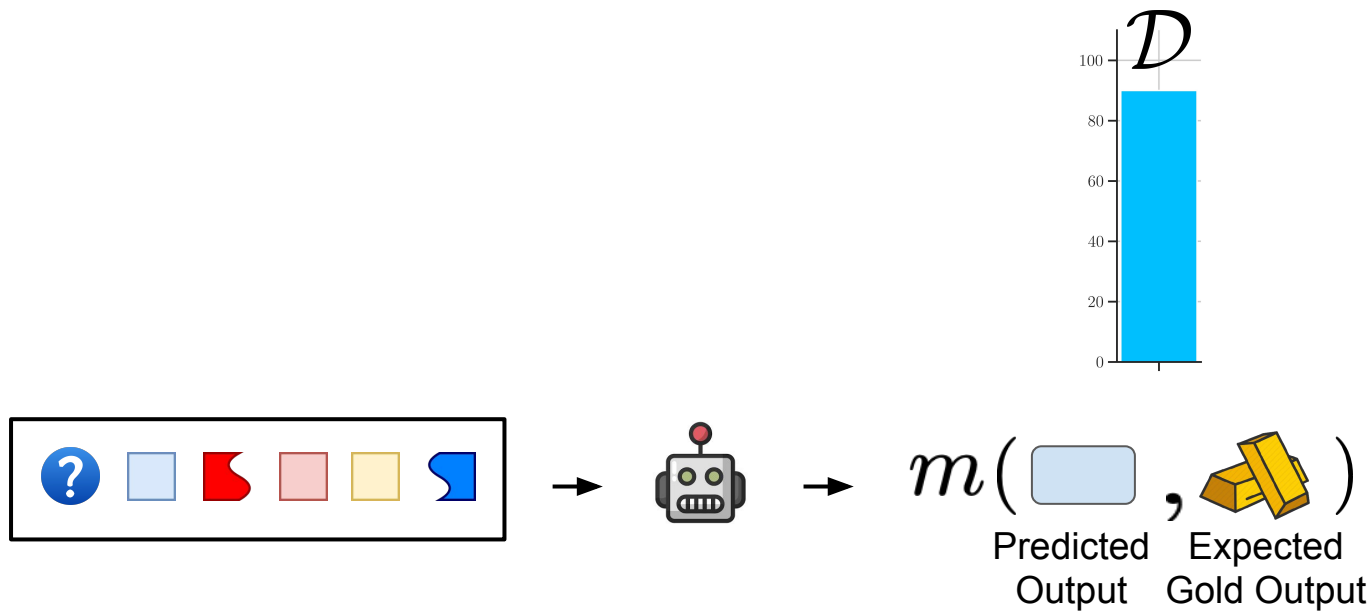
Measure Disconnected Reasoning of Model



DiRe Probing Dataset



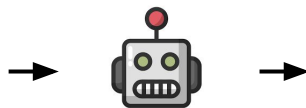
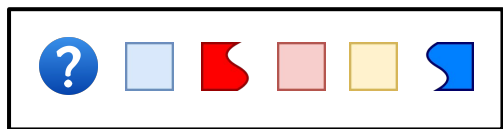
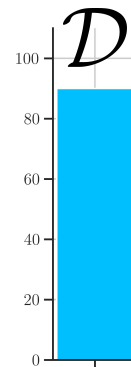
DiRe Probing Dataset $\mathcal{P}(\mathcal{D})$



Original Instance

DiRe Probing Dataset (\mathcal{D})

What extent of this metric m was achieved by disconnected reasoning?



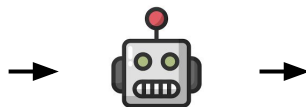
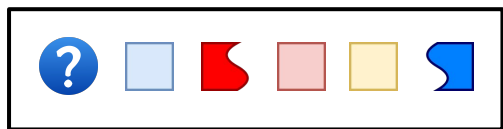
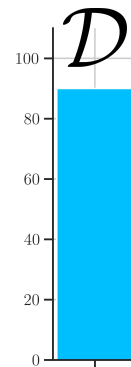
$m(\text{Predicted Output}, \text{Expected Gold Output})$

Original Instance

DiRe Probing Dataset (\mathcal{D})

What extent of this metric m was achieved by disconnected reasoning?

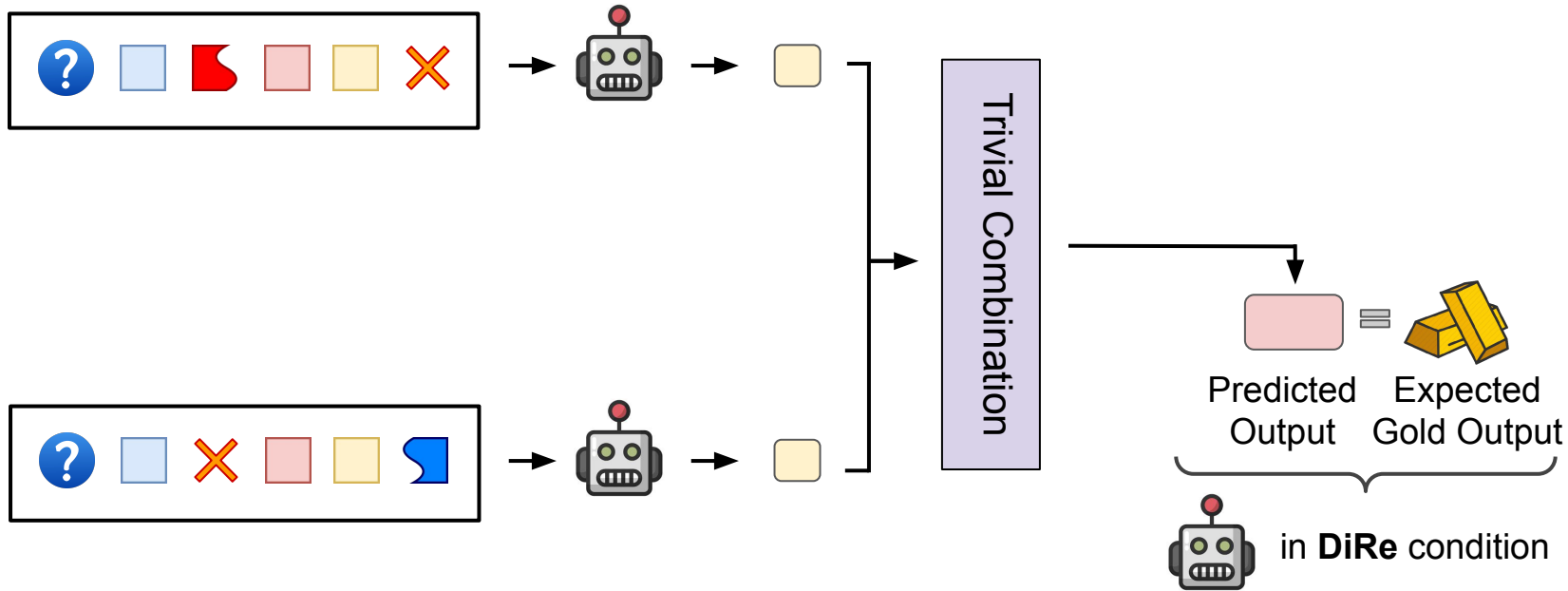
Use DiRe Condition



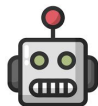
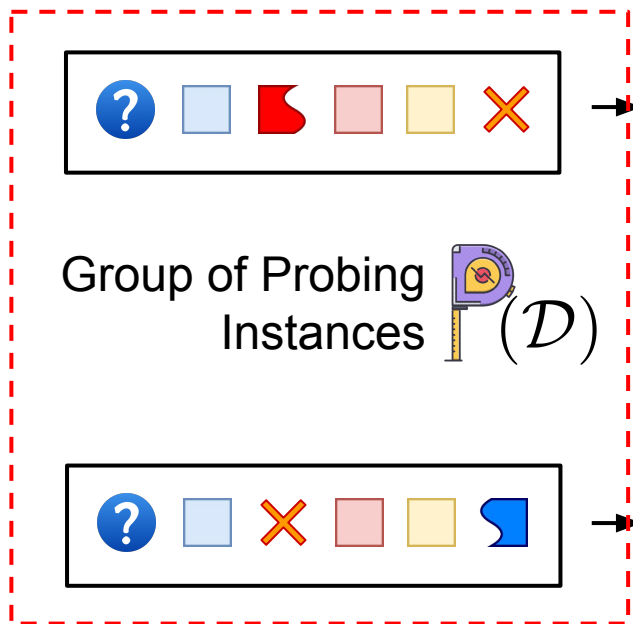
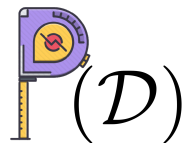
$$m(\text{Predicted Output}, \text{Expected Gold Output})$$

Original Instance

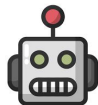
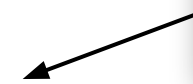
DiRe Probing Dataset (\mathcal{D})



DiRe Probing Dataset



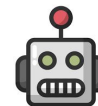
Create exhaustive partitions
such that no part contains all
supporting facts.



Combination

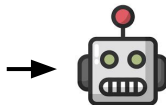
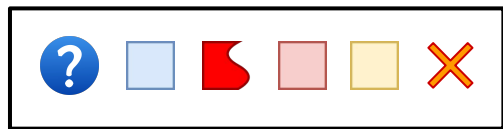
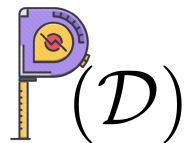


Predicted Output = Expected Gold Output



in **DiRe** condition

DiRe Probing Dataset



Group of Probing
Instances



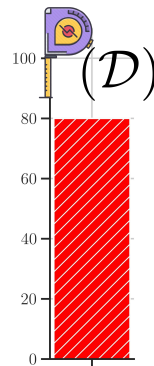
(\mathcal{D})



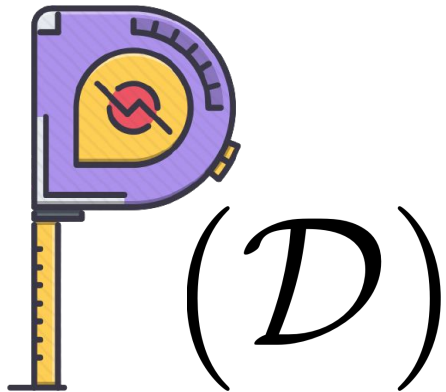
Compare trivially combined
output to gold output with the
metric that was used in \mathcal{D} .

Trivial
Combination

$m(\text{Predicted Output}, \text{Expected Gold Output})$




Measuring Disconnected Reasoning



DiRe **P**robing dataset of \mathcal{D}

Measure disconnected reasoning

- of model  on dataset \mathcal{D} .
- possible on the dataset \mathcal{D} .


Measure Disconnected Reasoning possible on \mathcal{D}

How far can we go on
 \mathcal{D} via disconnected reasoning?

Measure Disconnected Reasoning possible on \mathcal{D}

How far can we go on
 \mathcal{D} via disconnected reasoning?



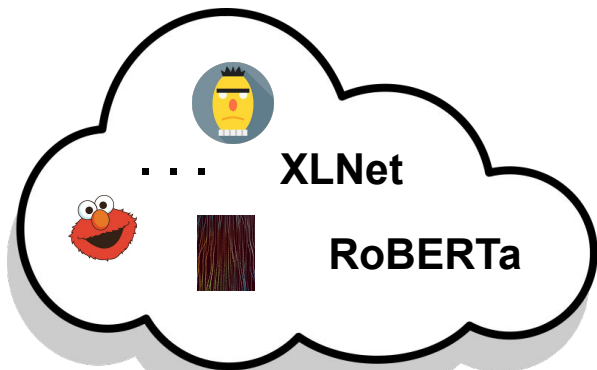
How far can we go on  (\mathcal{D}) ?

Measure Disconnected Reasoning possible on \mathcal{D}

How far can we go on
 \mathcal{D} via disconnected reasoning?



How far can we go on $\mathcal{P}(\mathcal{D})$?



Current state of NLP

Take best/strong
architecture and
optimize directly on



$$m(\bullet, \mathcal{P}(\mathcal{D}))$$

Details in the paper.

Outline of this work

We address these for reading comprehension multihop QA with annotated supporting facts.



- **Introduce Disconnected Reasoning (DiRe)**
 - a model-agnostic characterization of a form of non-multihop reasoning.

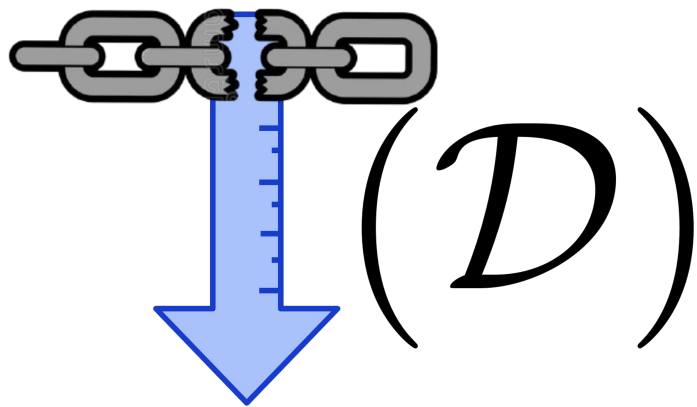


- **Measure** Disconnected Reasoning
 - Done by the given model.
 - Possible on the dataset.



- **Reduce** Disconnected Reasoning
 - Automatic transformation to reduce dataset cheatability.

Reducing Disconnected Reasoning

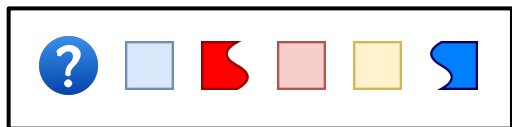


Transform the dataset *automatically* to **reduce** its cheatability via **disconnected reasoning**.


Support Sufficiency

Support Sufficiency

Answerable question with
Sufficient supporting facts



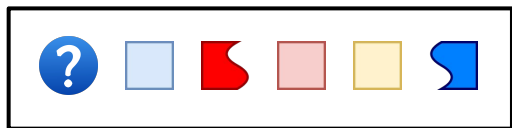

→
Disconnected
Reasoning


Answer	SF
India	 

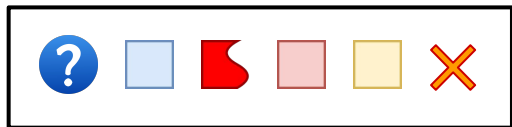
Doesn't use information of one
supporting fact for the selection
and use of the other.

Support Sufficiency

Answerable question with
Sufficient supporting facts




Drop 
to make question
unanswerable



Unanswerable question with
insufficient supporting facts


→
Disconnected
Reasoning

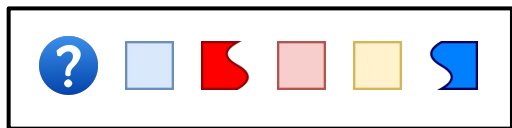
Answer	SF
India	 


→
Disconnected
Reasoning

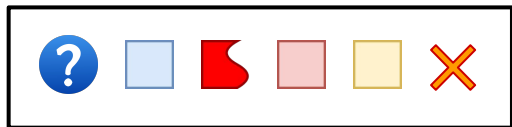
Answer	SF
India	 

Support Sufficiency

Answerable question with
Sufficient supporting facts




Drop 
to make question
unanswerable




Unanswerable question with
insufficient supporting facts


→
Disconnected
Reasoning



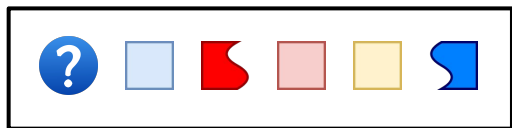

→
Disconnected
Reasoning



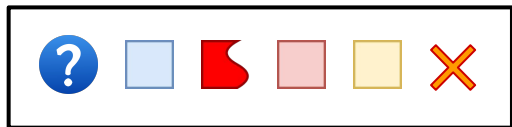
 **doesn't check**
if the supporting
facts **connect**, so
cannot easily
distinguish these
cases.

Support Sufficiency

Answerable question with
Sufficient supporting facts



Drop 
to make question
unanswerable

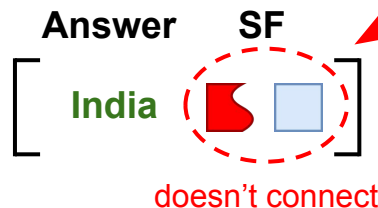



Unanswerable question with
insufficient supporting facts


→
Disconnected
Reasoning




→
Disconnected
Reasoning



 **checks** if the
supporting facts
connect, so **can**
easily **distinguish**
these cases.

Support Sufficiency

Answerable question with
Sufficient supporting facts



Answer

SF



How to use this idea to make the dataset \mathcal{D}
harder for disconnected reasoning?

Checks if the
ing facts
, so **can**
stinguish
ses.



Disconnected
Reasoning

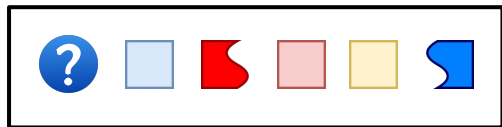
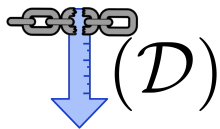
India



doesn't connect

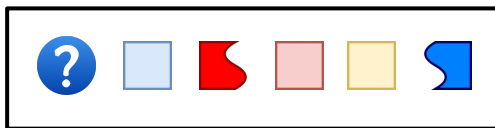
Unanswerable question with
insufficient supporting facts

Transformed Dataset

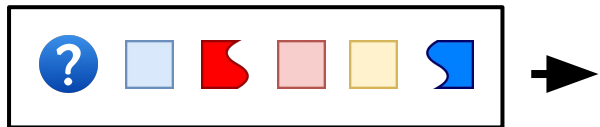


Original Instance

Transformed Dataset  (\mathcal{D})



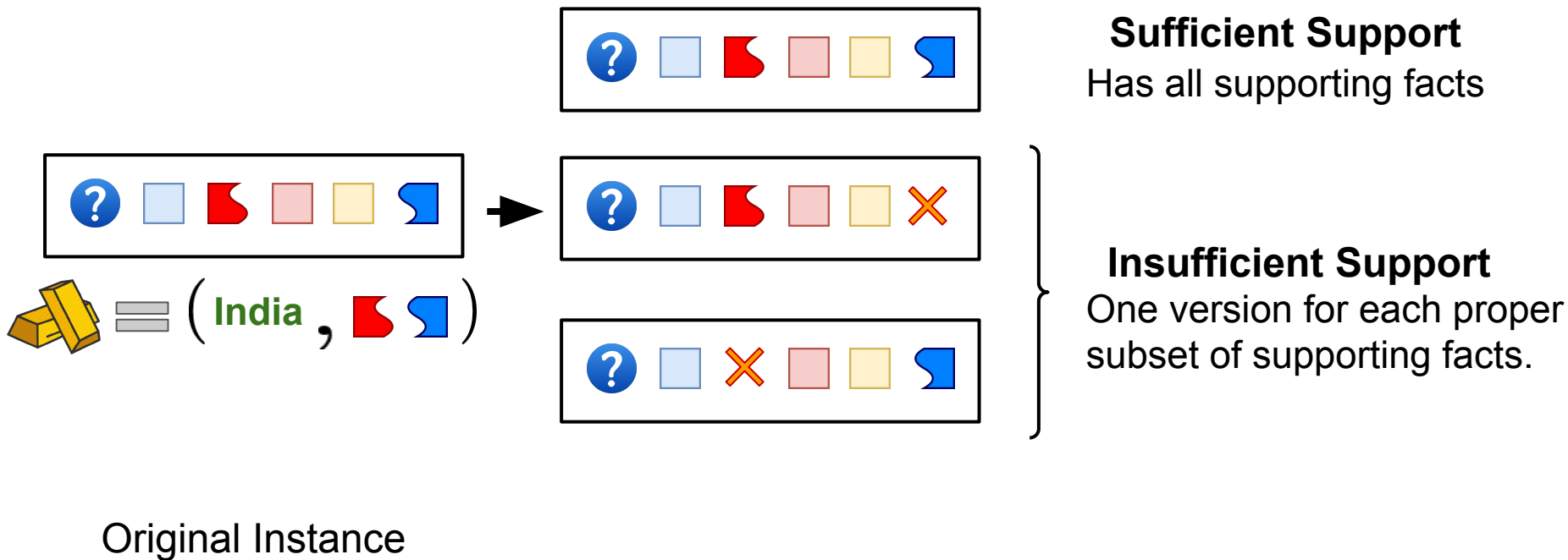
Sufficient Support
Has all supporting facts



 = (India,  )

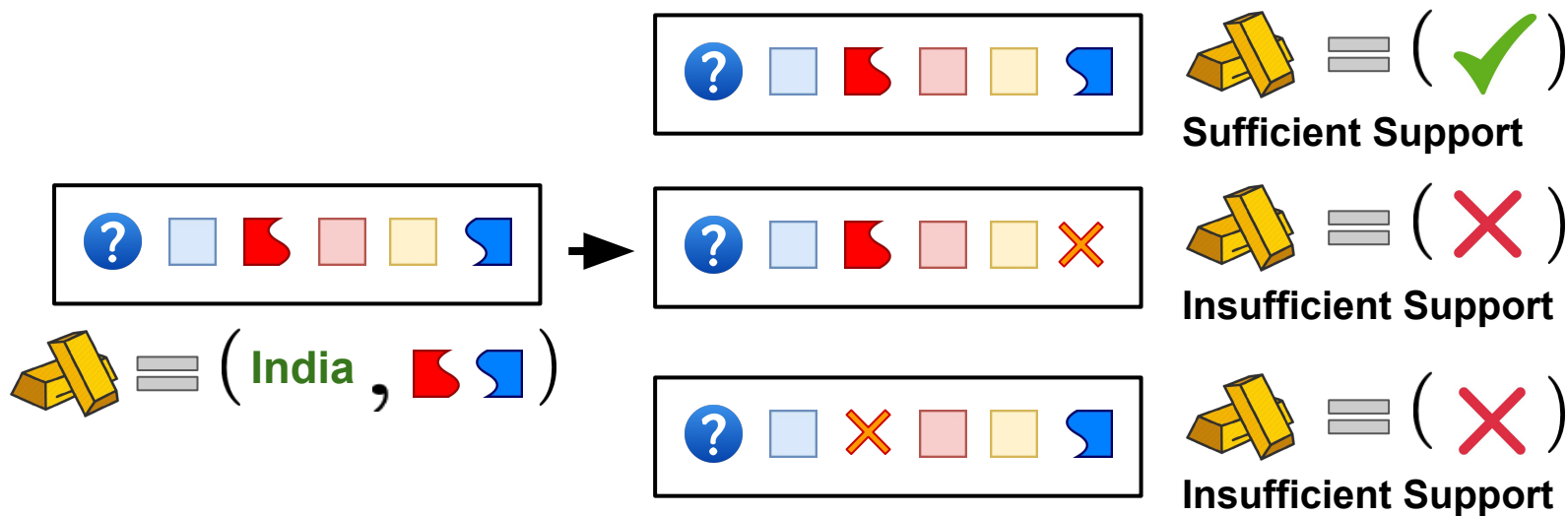
Original Instance

Transformed Dataset (\mathcal{D})



Transformed Dataset (\mathcal{D})

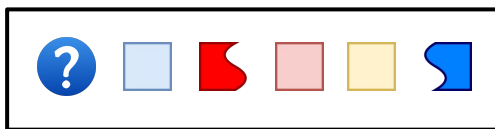
Contrastive Support Sufficiency (CSS) Test



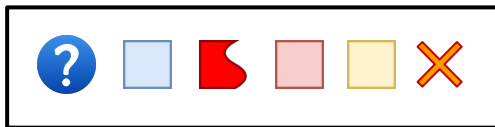
Transformed Dataset (\mathcal{D})



CSS + Answer + Supporting Facts Tests

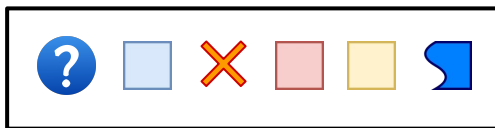
CSS Test can be combined with **Answer** and **Supporting Facts** tests, by simply adding the gold labels to sufficient support instance.





 = (, India,  )
Sufficient Support



 = ()
Insufficient Support

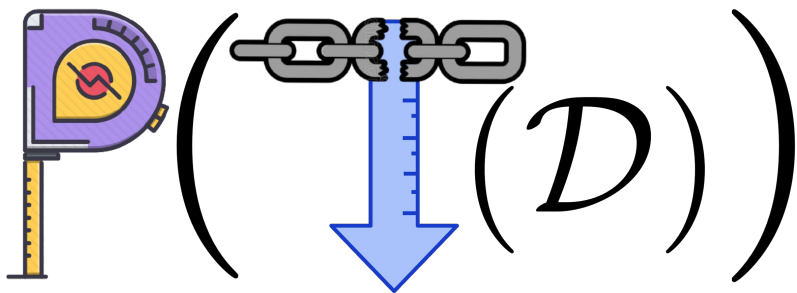


 = ()
Insufficient Support

Transformed Dataset Instances

How cheatable is the Transformed Dataset?

How cheatable is the Transformed Dataset?



DiRe Probing dataset of
the Transformed dataset

- We can go back to DiRe condition again, and derive its DiRe probing dataset $P(T(\mathcal{D}))$.
- We'll show measurements on this probe but details are in paper.

Experiment Setup

Dataset (\mathcal{D})

- Multihop RC dataset: **HotpotQA** (113K questions).
- Each question has a set of **10** wikipedia paragraphs as context, with **2** supporting paragraphs and few sentences.
- We have answer (**Ans**), paragraph support (**Supp_p**) and sentence support (**Supp_s**) metrics their combinations.

Models ()

- RNN Baseline
- XLNet-base

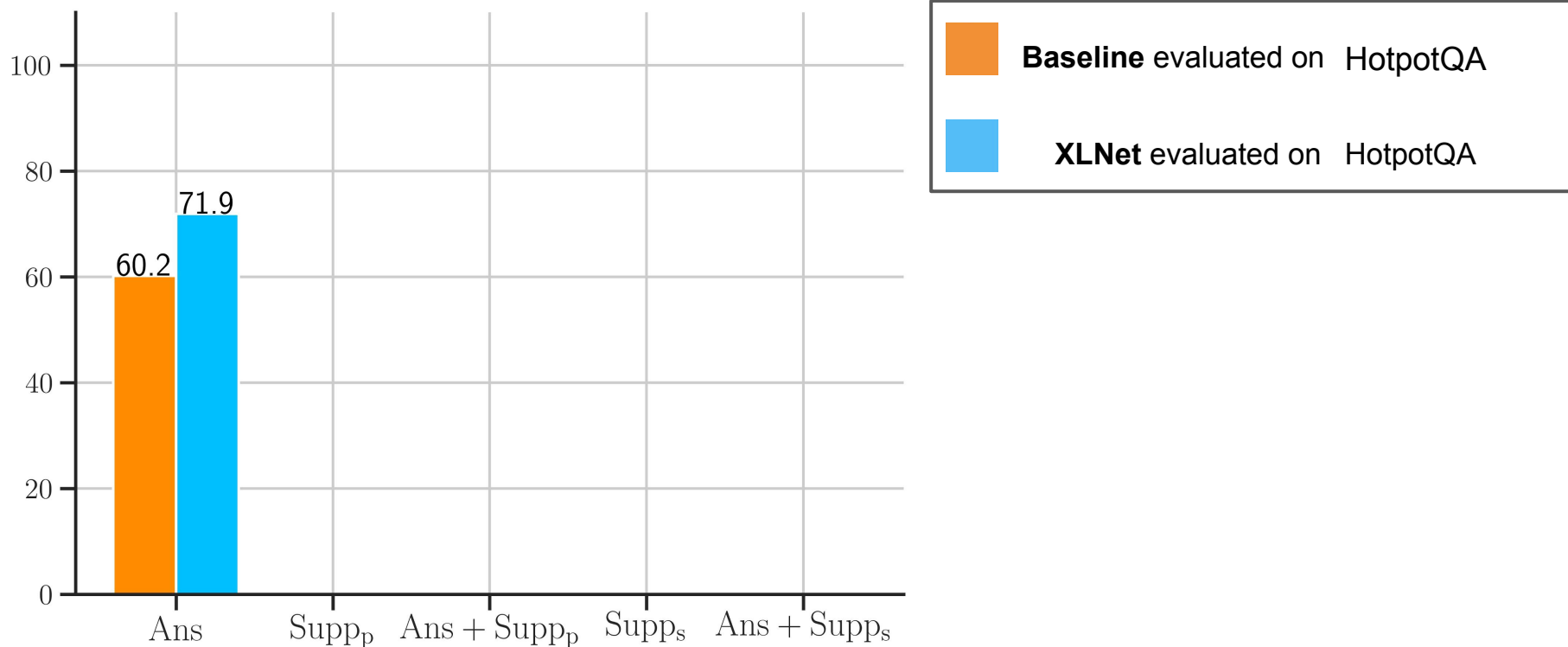
Evaluation

Q1 How much disconnected reasoning models do on HotpotQA?

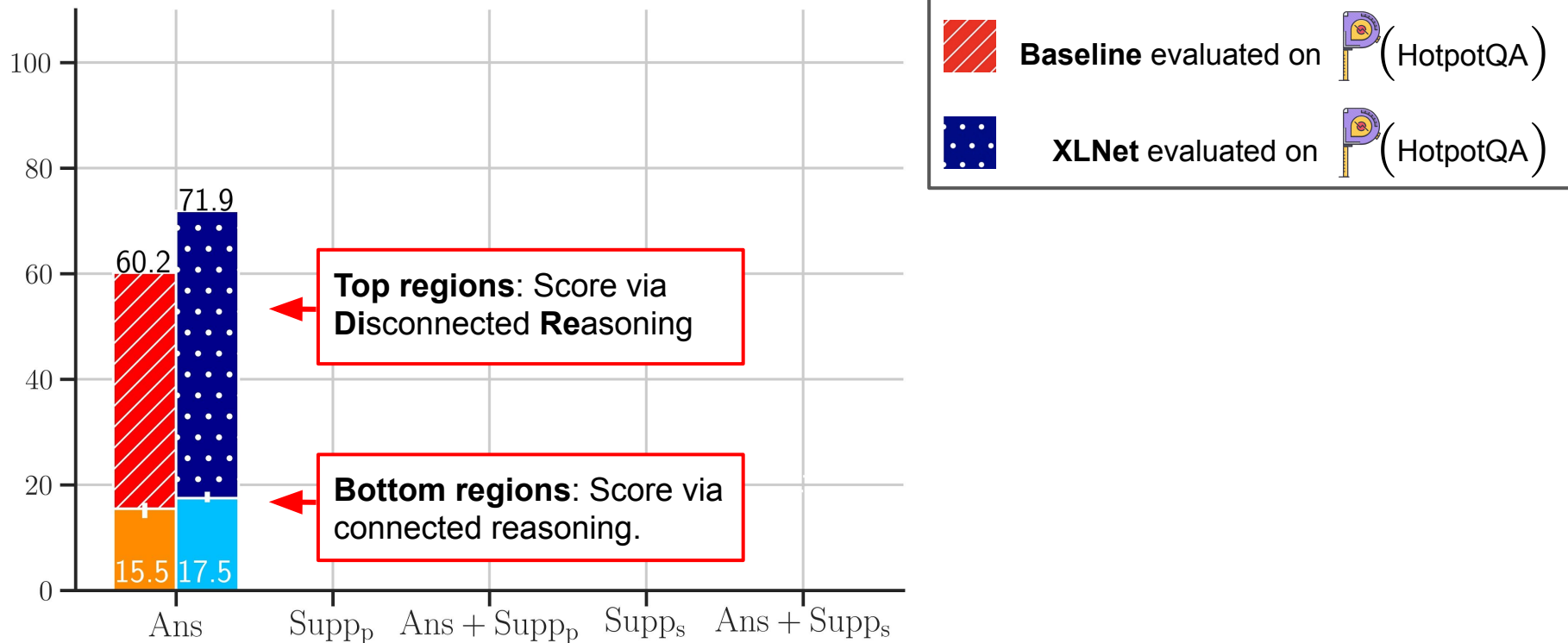
Q2 Does the transformation reduce DiRe cheatability of HotpotQA?

Q1 How much disconnected reasoning models do on HotpotQA?

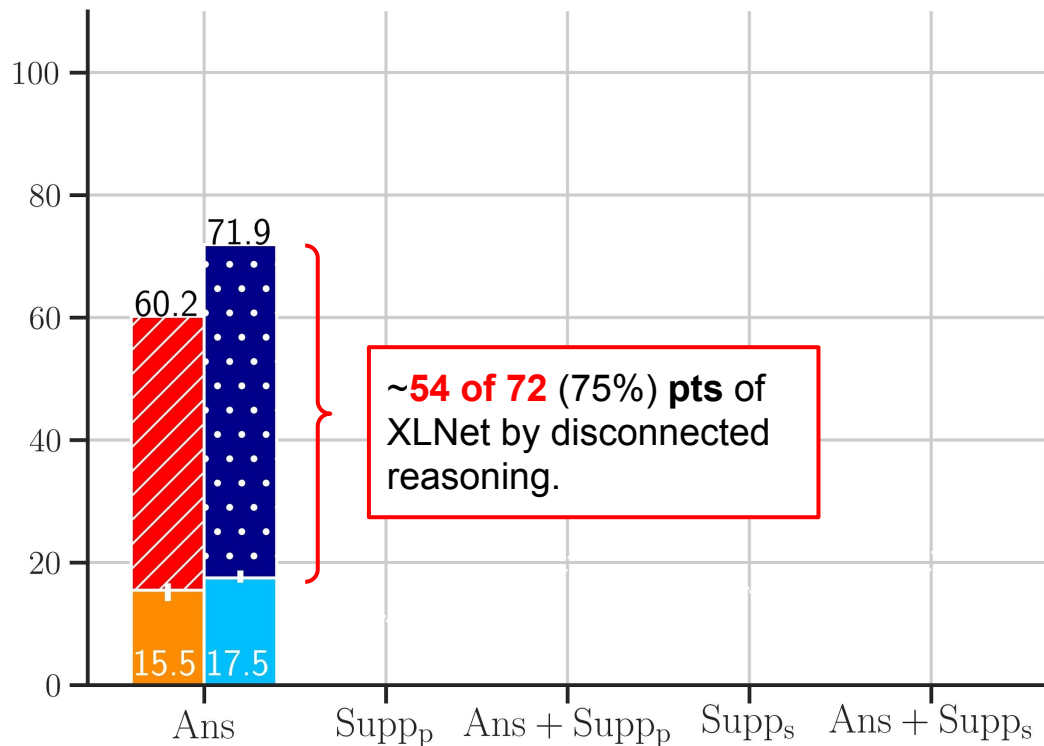
Q1 How much disconnected reasoning models do on HotpotQA?







Q1 How much disconnected reasoning models do on HotpotQA?



Q1 How much disconnected reasoning models do on HotpotQA?

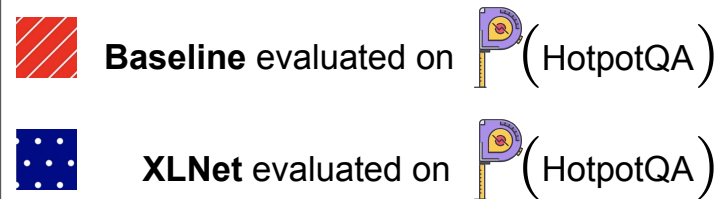
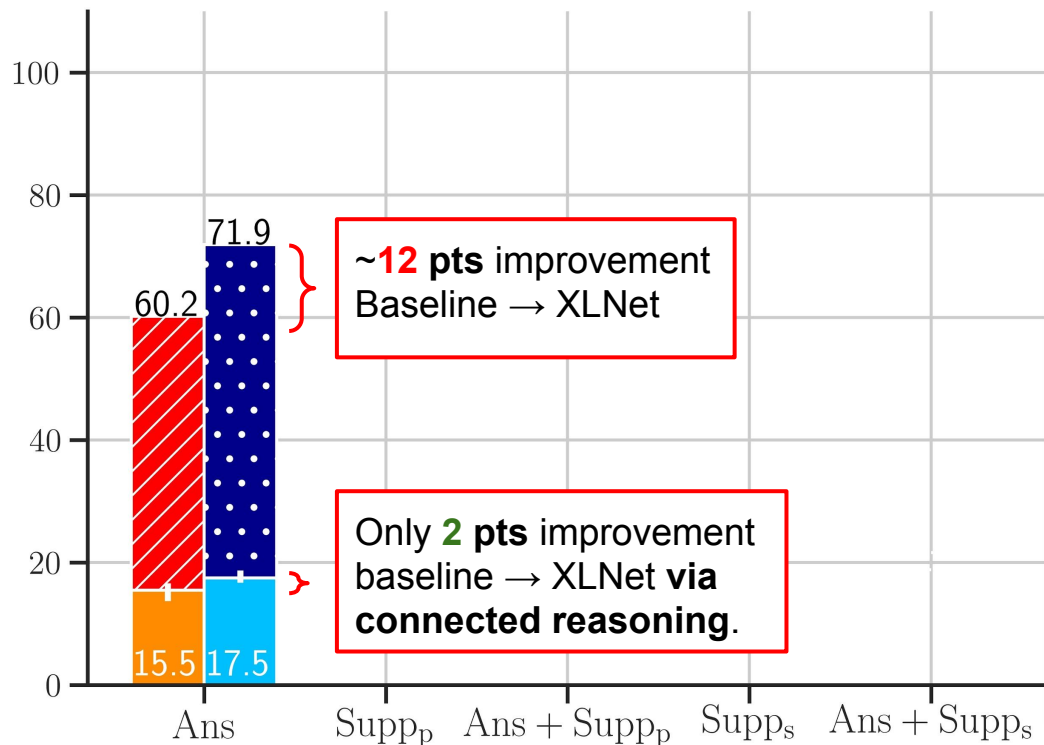


 **Baseline** evaluated on  (HotpotQA)

 **XLNet** evaluated on  (HotpotQA)

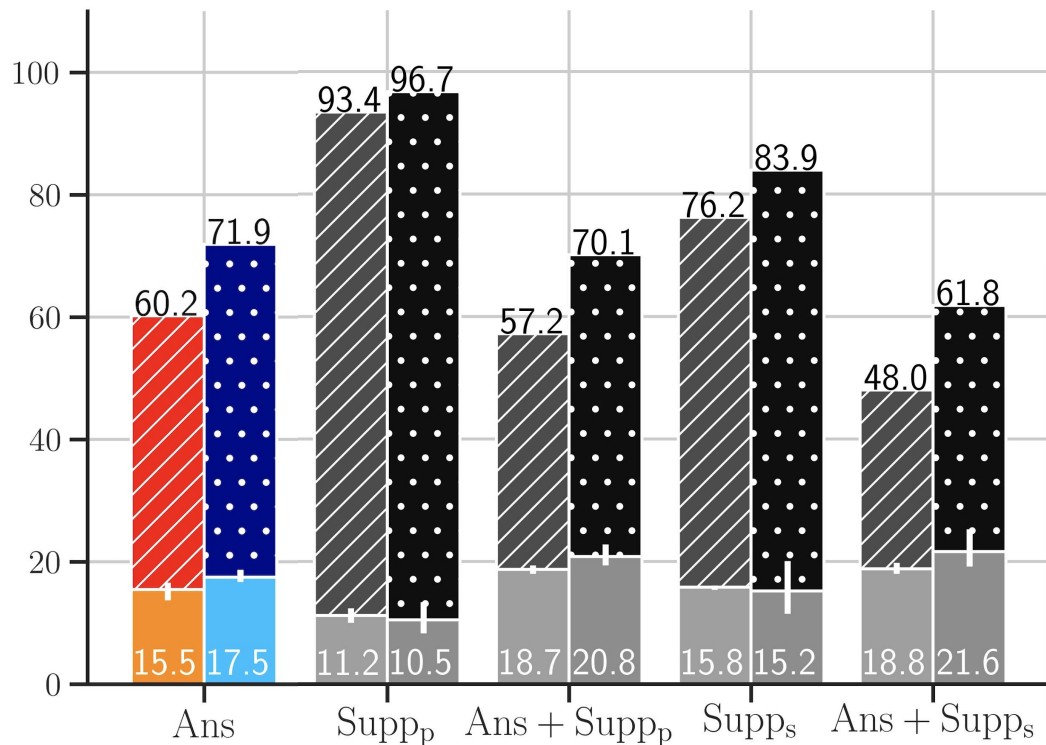
Most of the total model scores can be attributed to disconnected reasoning.



Q1 How much disconnected reasoning models do on HotpotQA?





- Although there's large total score improvement from Baseline to XLNet,
- Only a small amount (**2 pts**) is through connected reasoning.

Q1 How much disconnected reasoning models do on HotpotQA?



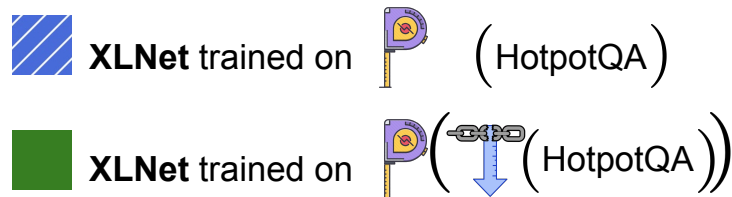
 **Baseline** evaluated on  (HotpotQA)

 **XLNet** evaluated on  (HotpotQA)

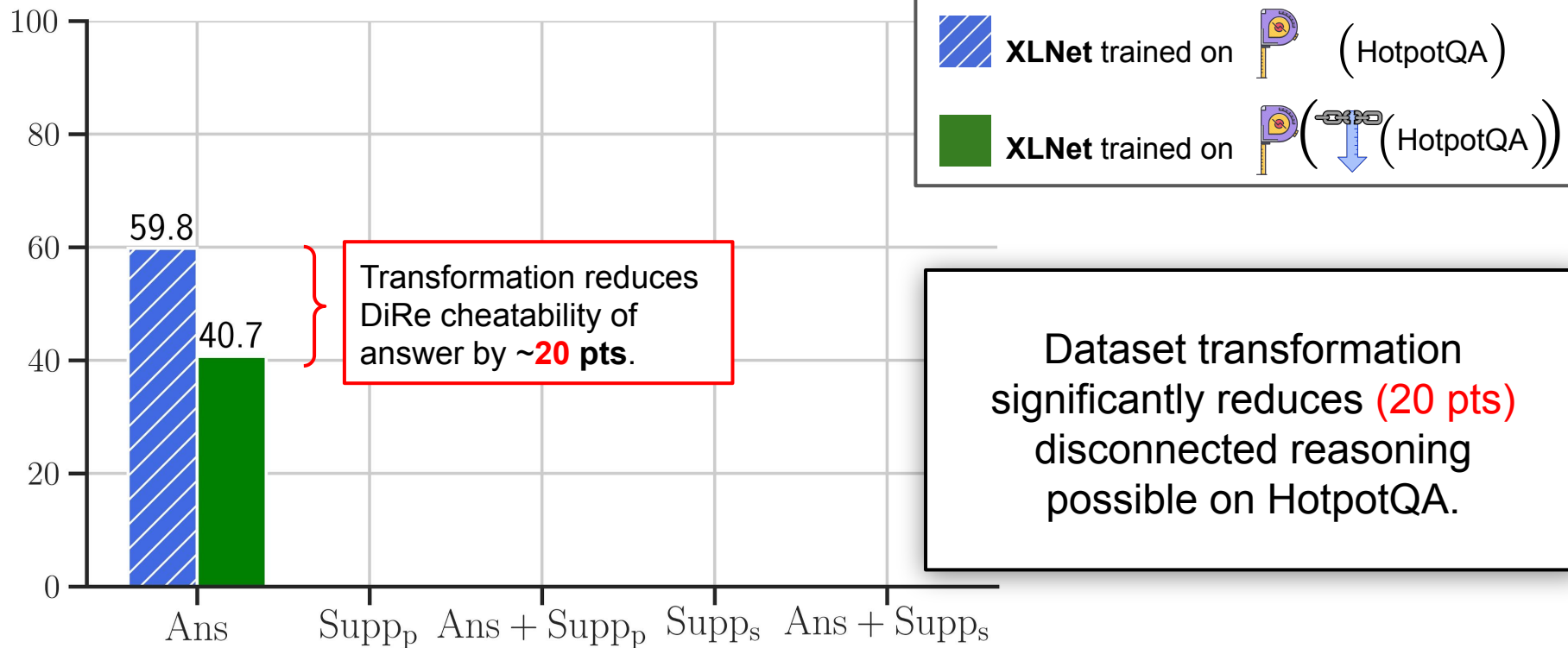
- Although there's large total score improvement from Baseline to XLNet,
- Only a small amount (2 pts) is through connected reasoning.

Q2 Does the transformation reduce DiRe cheatability of HotpotQA?

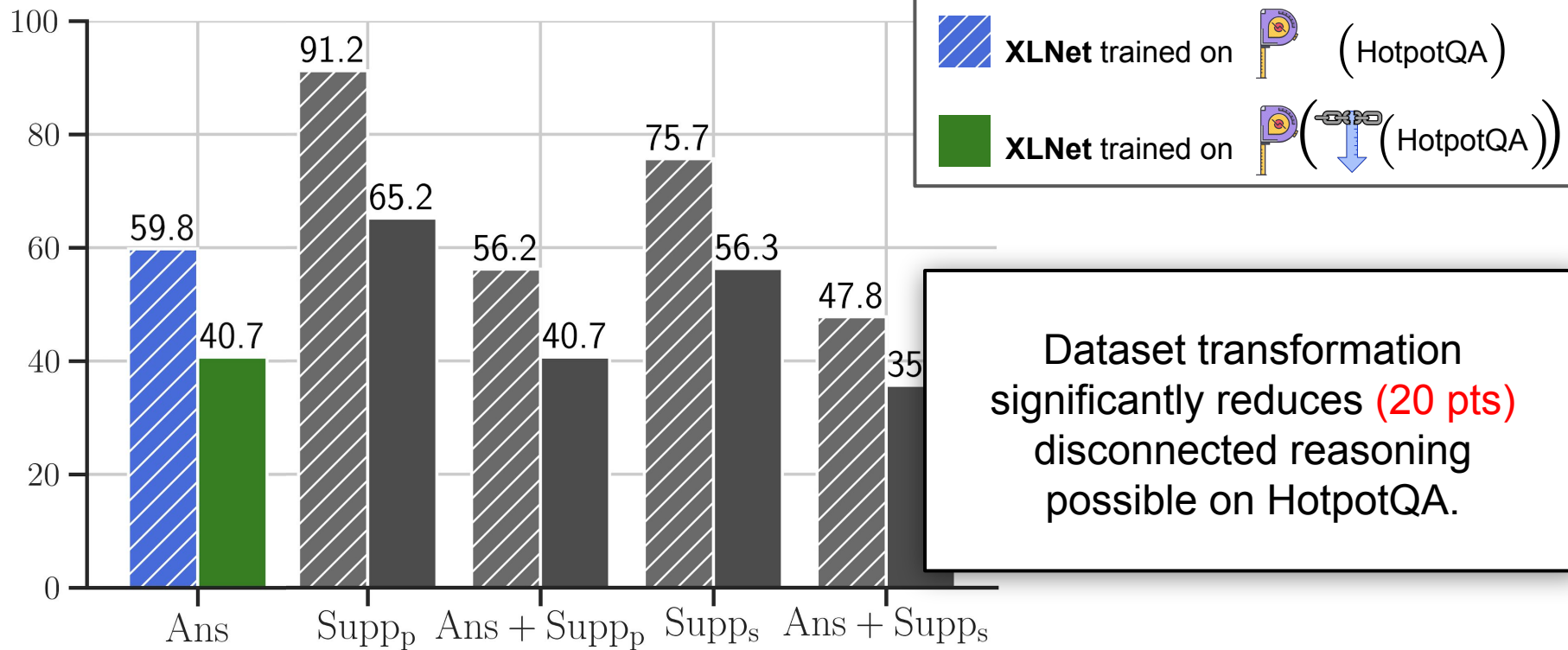
Q2 Does the transformation reduce DiRe cheatability of HotpotQA?



Q2 Does the transformation reduce DiRe cheatability of HotpotQA?



Q2 Does the transformation reduce DiRe cheatability of HotpotQA?



Other Experiments (in paper)

Our transformation  :

- Makes the dataset significantly harder for models, while being not much harder for humans.
- Is significantly more effective than, and complementary to, adversarial method of removing such biases.

Conclusion

- Introduced **Disconnected Reasoning**, a form of undesirable reasoning prevalent in multihop models, and devised model-agnostic probe to catch such behavior.
- Showed large portion of progress in multifact reasoning can be attributed to disconnected reasoning.
- Introduced *Contrastive Support Sufficiency* to make existing support-annotated multi-hop datasets more difficult and less cheatable.

Takeaway Tools



Takeaway Tools



Multihop **Model**
Designer

can use



Measure DiRe cheating
of your model.

Takeaway Tools



Multihop **Model**
Designer

can use



Measure DiRe cheating
of your model.

Multihop **Data**
Designer

can use



Measure DiRe cheatability
of your dataset.

Takeaway Tools



Multihop **Model**
Designer

can use



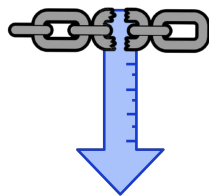
Measure DiRe cheating
of your model.

Multihop **Data**
Designer

can use



Measure DiRe cheatability
of your dataset.



Reduce DiRe cheatability
of your dataset.

Takeaway Tools

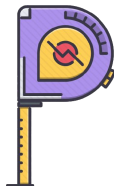


Code Available:

<https://github.com/stonybrooknlp/dire>

Multihop **Model**
Designer

can use



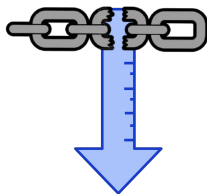
Measure DiRe cheating
of your model.

Multihop **Data**
Designer

can use



Measure DiRe cheatability
of your dataset.



Reduce DiRe cheatability
of your dataset.