Partial-input baselines show that NLI models can ignore context, but they don't.

Neha Srikanth & Rachel Rudinger
Department of Computer Science, University of Maryland College Park

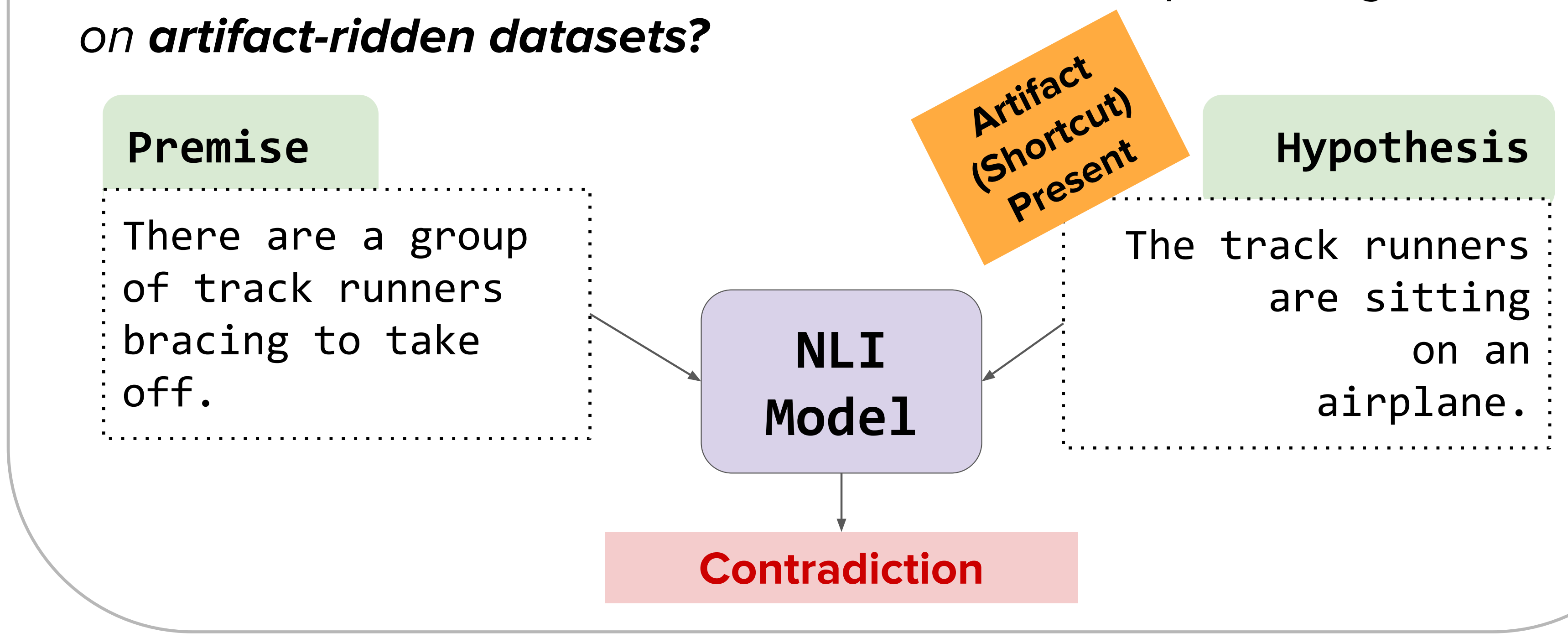
SCAN ME

Motivation

A necessary, but not sufficient, condition of true inferential reasoning is the ability for NLI models to **utilize all parts of the example's input.**

Many claim that datasets containing annotation artifacts may produce models incapable of learning to perform such reasoning.

*Do NLI models **learn to condition on context** despite being trained on **artifact-ridden datasets?***


Premise

There are a group of track runners bracing to take off.

Artifact (Shortcut) Present

Hypothesis

The track runners are sitting on an airplane.

NLI Model

Contradiction

NLI Datasets

SNLI (Bowman et. al 2015)

Determine whether **premise (P)** *entails*, *contradicts*, or is *neutral* with respect to a **hypothesis (H).**

Context

`Premise: A man is sitting in a dim restaurant.`

Target

`Hypothesis: He is eating food.`

`Label: Neutral`

***δ*-NLI (Rudinger et. al 2020)**

When H **is neutral,** determine whether a third **update (U)** sentence *strengthens* or *weakens* H.

Context

`Premise: A man is sitting in a dim restaurant.`
`Hypothesis: He is eating food.`

Target

`Update: He is browsing a menu.`

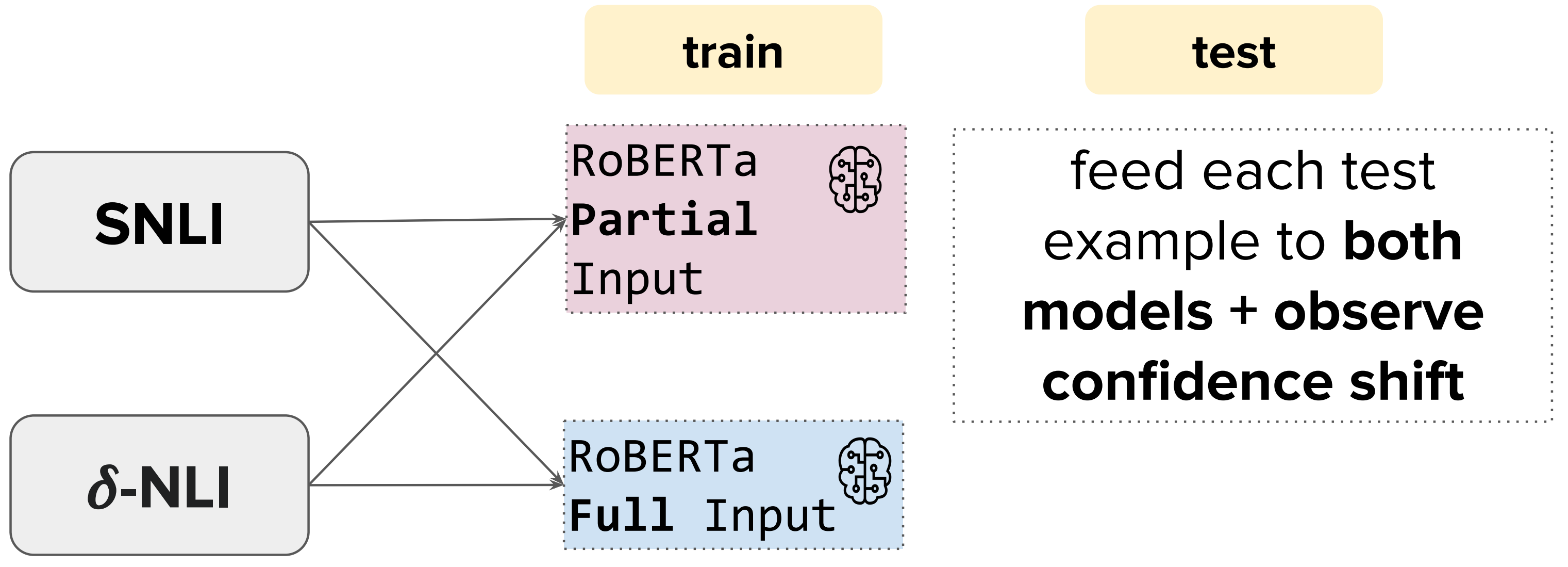
`Label: Weakener`

Experiment 1: Context in NLI

Strong partial-input models demonstrate that **full-input models do not necessarily need to utilize context** to make correct predictions.

But do they?

And, for examples which partial-input baselines predict the label correctly, ***does access to context shift a full-input model's confidence in the correct label?***


train

test

SNLI

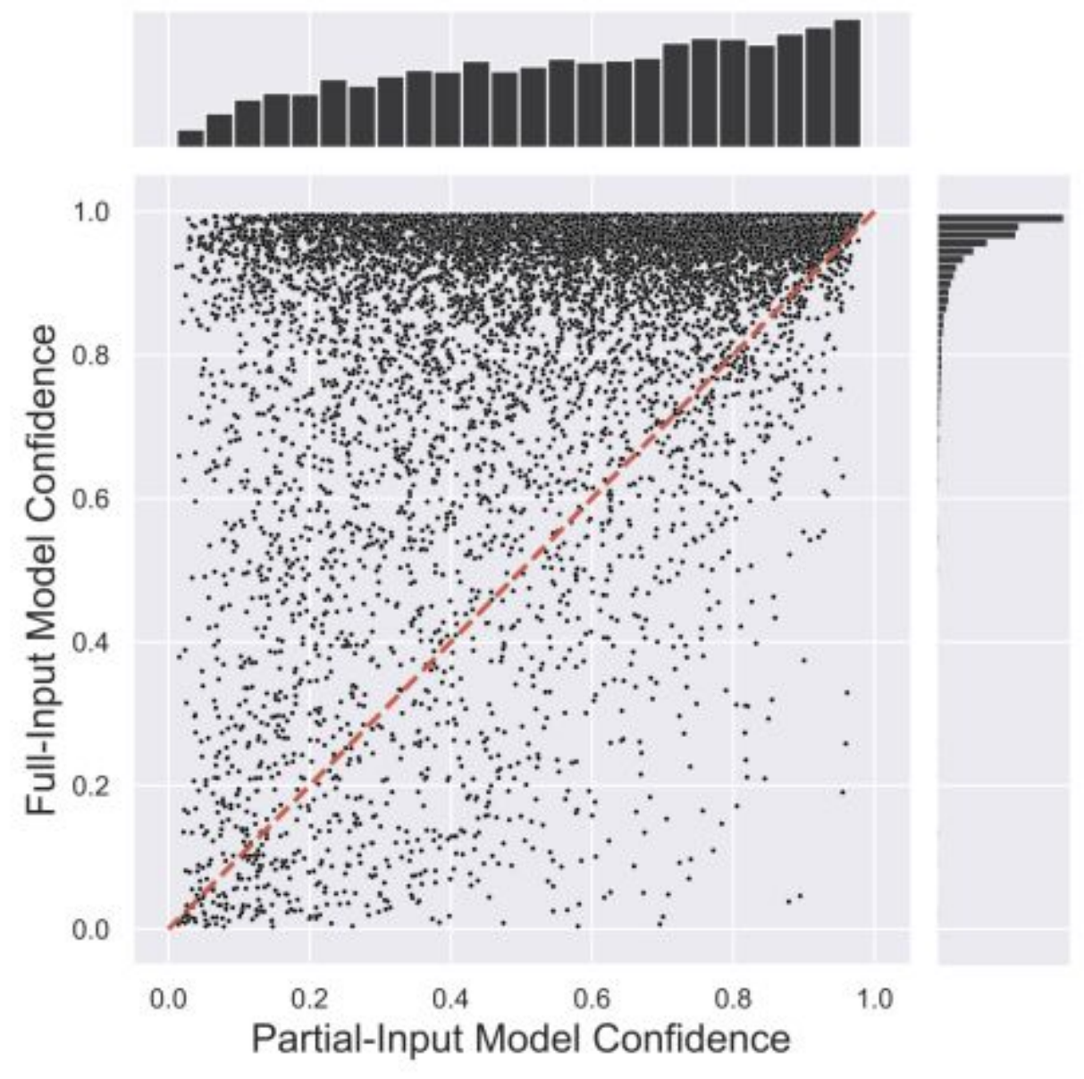
δ-NLI

RoBERTa Partial Input

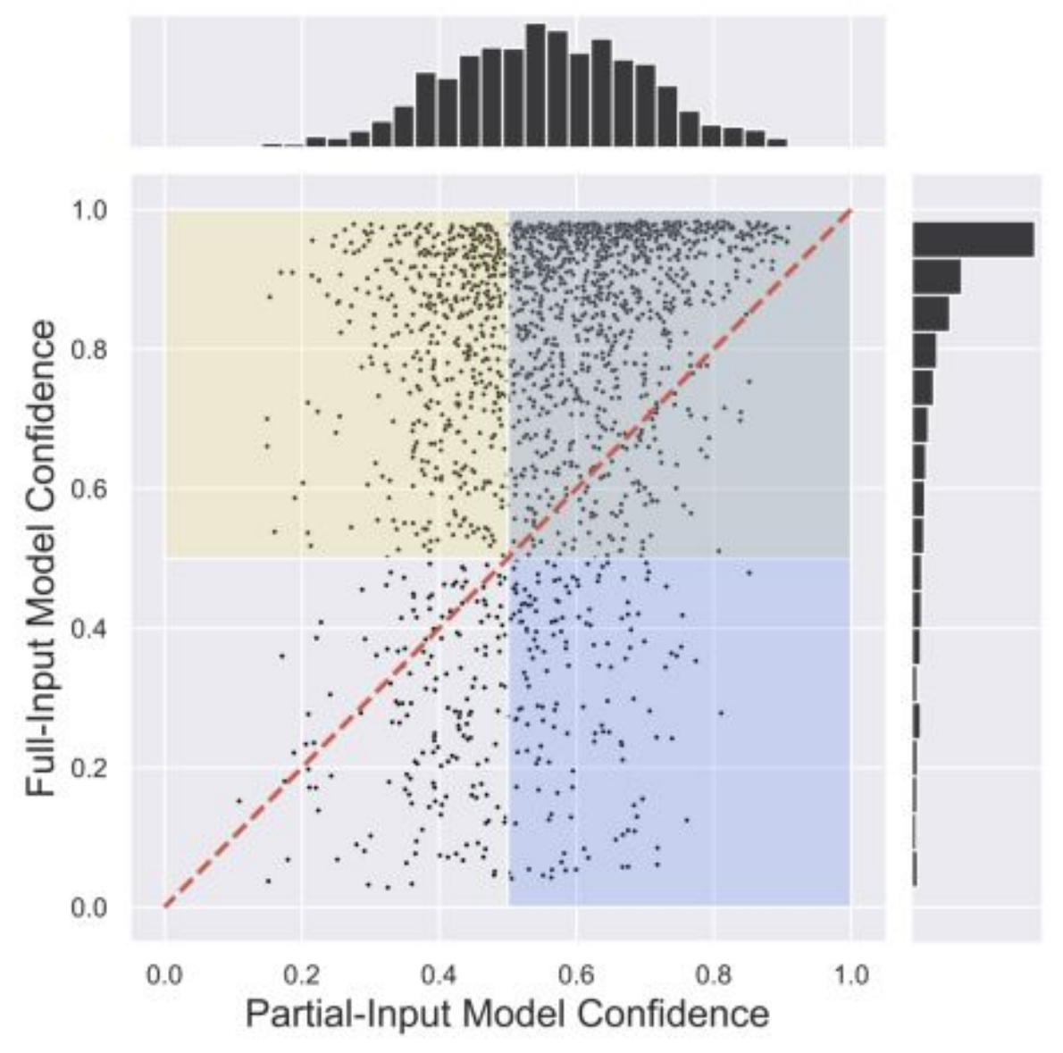
RoBERTa Full Input

feed each test example to **both models** + **observe confidence shift**

SNLI Confidence Shifts



***δ*-NLI Confidence Shifts**



Plots of ordered pairs of each model's confidence in the **correct label** for test examples (partial-input along the x-axis & full-input along the y-axis).

Density around the diagonal would indicate no change in confidence.

As evidenced by **density above the diagonal, full-input models are more confident in the correct label.** This behavior hints that **full-input models may be successfully learning to leverage additional context** instead of overgeneralizing on artifacts in the target.

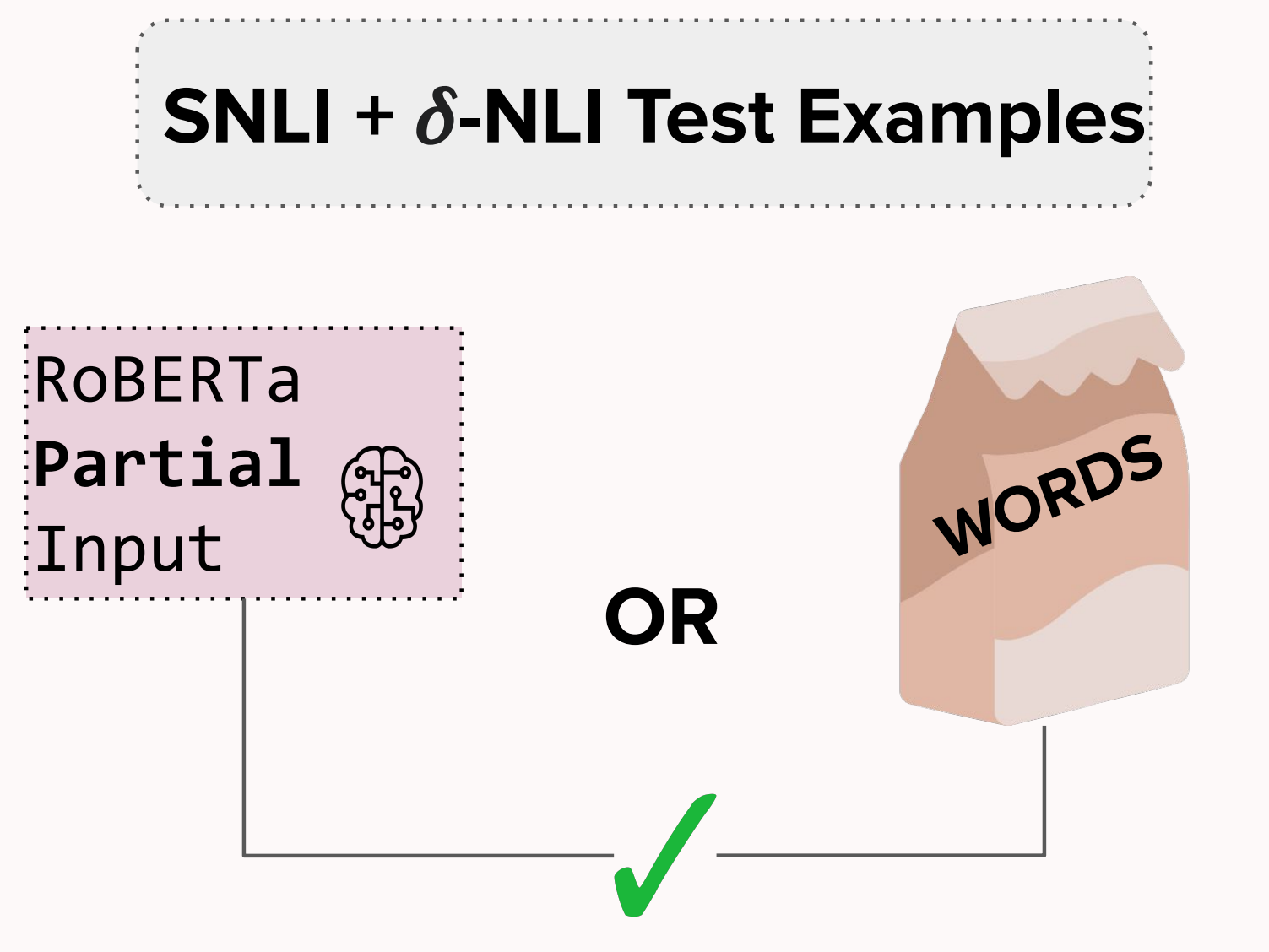
Experiment 2: Context Editing

We investigate a model's ability to leverage context despite artifacts by exploring **how sensitive full-input models are to changes in non-target components of the input.**

We present an example modification scheme in which **we edit context sentences from examples where a model correctly predicts the label from the target alone**. Our final evaluation set consists of 600 examples sourced from SNLI and *δ*-NLI.

1 Example Subselection

Identify examples **most likely to contain artifacts**

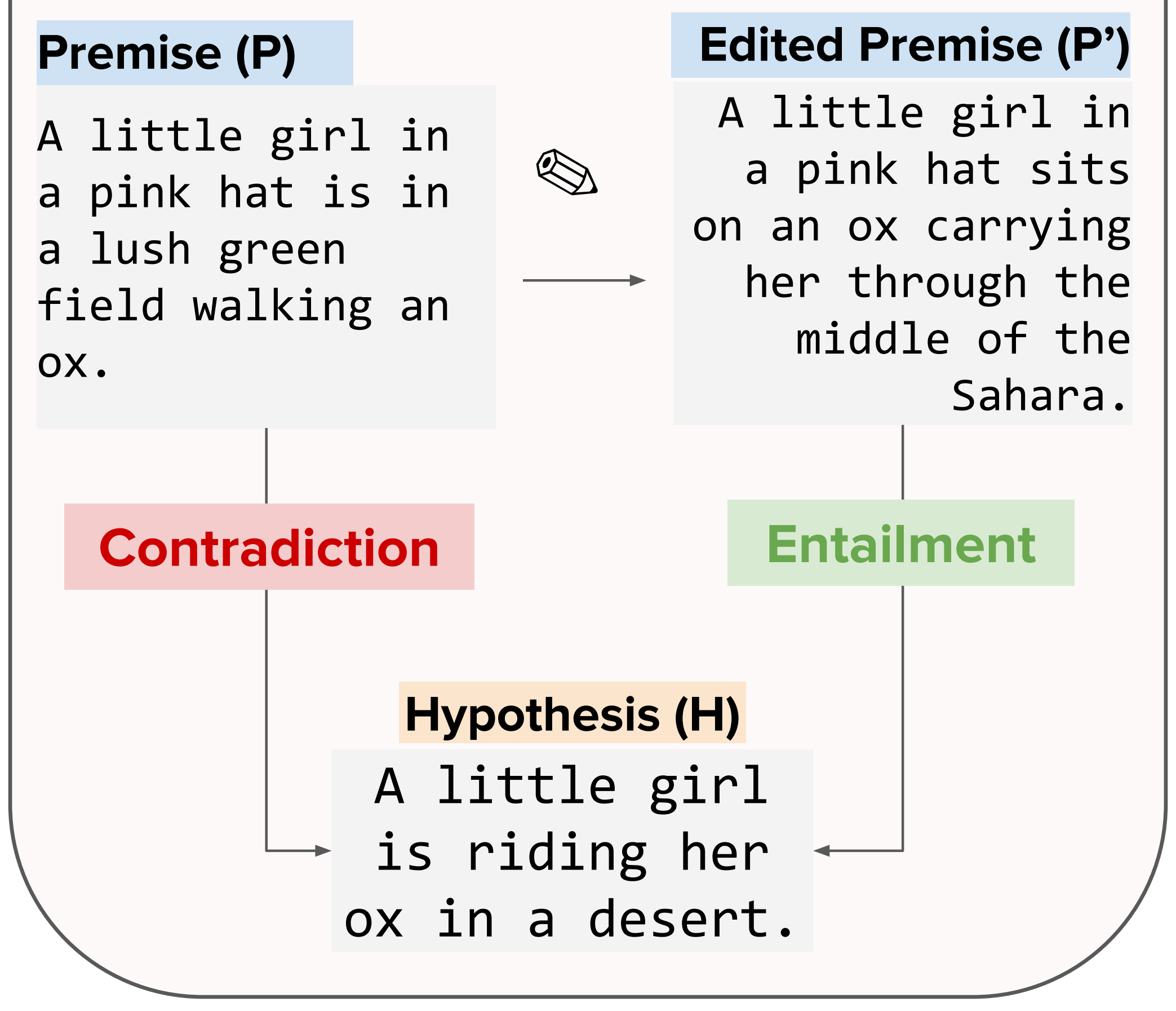

SNLI + *δ*-NLI Test Examples

RoBERTa Partial Input

OR

✓

2 Example Editing


Premise (P)

`A little girl in a pink hat is in a lush green field walking an ox.`

Edited Premise (P')

`A little girl in a pink hat sits on an ox carrying her through the middle of the Sahara.`

Contradiction

Entailment

Hypothesis (H)

`A little girl is riding her ox in a desert.`

3 Construct Expert-Annotated & Validated Evaluation Set

Final evaluation set consists of **600 examples containing edited context-target pairs split evenly across SNLI and *δ*-NLI** and balanced across original and edited target labels.

Agreement (Cohen's κ)	
SNLI	***δ*-NLI**
0.78	0.76

All examples were manually edited by one author and **independently validated** by another.

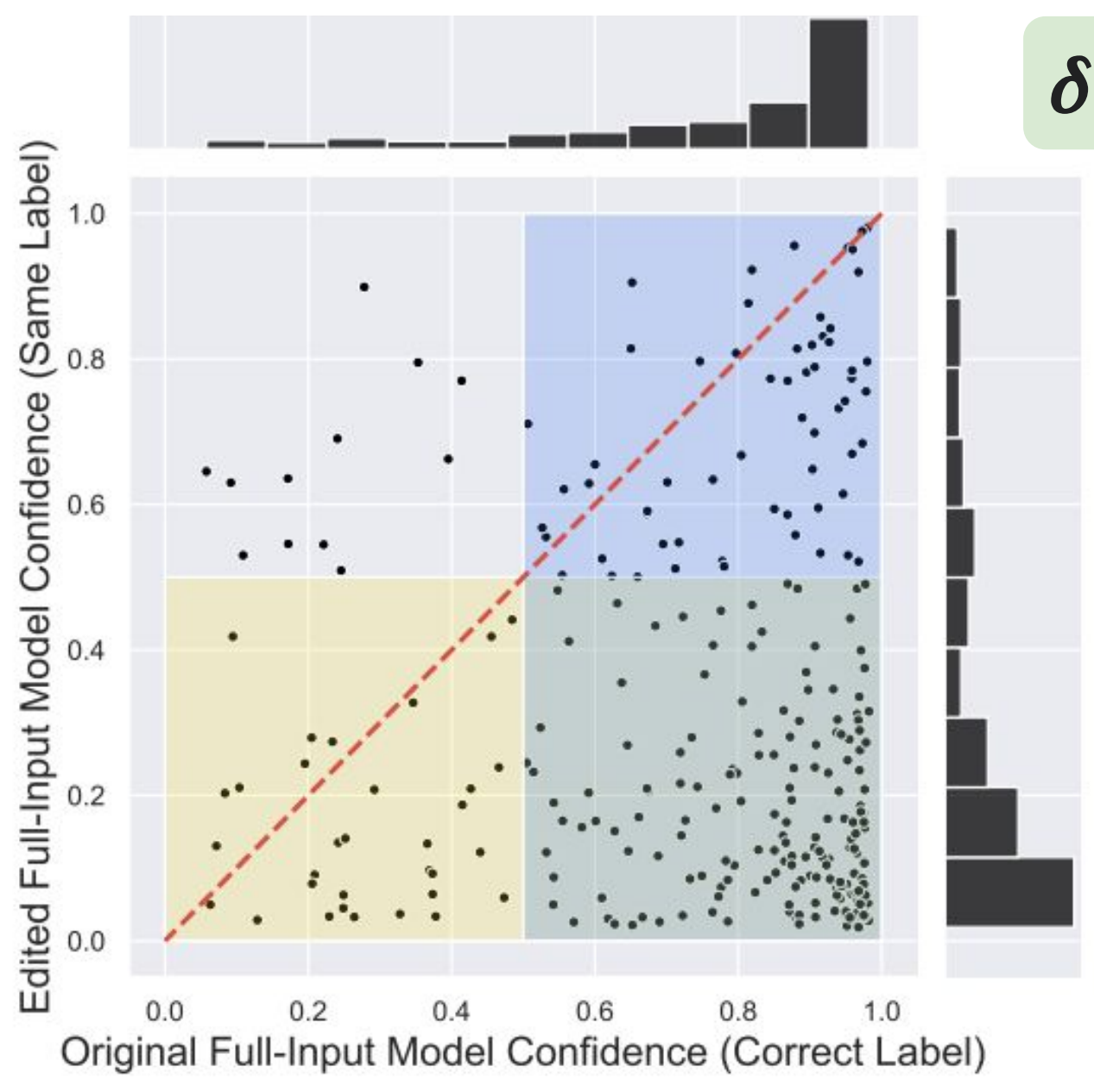
SNLI

		Edited Label (*l'*)		
		e	**n**	**c**
Original Label (*l*)	**e**	-	0.76	0.76
	n	0.42	-	0.78
	c	0.90	0.78	-

***δ*-NLI**

		Edited Label (*l'*)	
		w	**s**
Original Label (*l*)	**w**	-	0.75
	s	0.72	-

***δ*-NLI**


Edited Full-Input Model Confidence (Same Label)
Original Full-Input Model Confidence (Correct Label)