Systematic Probing for Examining Evidence Based Tabular Reasoning

Vivek Gupta[1], Riyaz Bhat[2], Atreya Ghosal[2],Manish Shrivastava[3], Maneesh Singh[3], Vivek Srikumar[1]

[1]University of Utah; [2]Verisk Inc.; [3]IIIT-Hyderabad

1. Tabular Inference Problem

- Inference task where premises are tabular in nature
- Given a premise table determine if a hypothesis is true (entailment), false (contradiction), or undetermined (neutral).

New York Stock Exchange	
Type	Stock exchange
Location	New York City, New York, U.S.
Founded	May 17, 1792; 226 years ago
Currency	United States dollar
No. of listings	2,400
Volume	US$20.161 trillion (2011)

H1: NYSE has fewer than 3,000 stocks listed.
H2: Over 2,500 stocks are listed in the NYSE.
H3: S&P 500 stock trading volume is over $10 trillion.

- Example InfoTabS dataset (Gupta et al., 2020), **H1**: entailed ; **H2**: contradictory ; **H3**: neutral

2. Tabular Reasoning

Evidence Based Reasoning

New York Stock Exchange	
Type	Stock exchange
Location	New York City, New York, U.S.
Founded	May 17, 1792; 226 years ago
Currency	United States dollar
No. of listings	2,400
Volume	US$20.161 trillion (2011)

H1: NYSE has fewer than 3,000 stocks listed.
H2: Over 2,500 stocks are listed in the NYSE.
H3: S&P 500 stock trading volume is over $10 trillion.

The *row* **No. of Listing** is required to establish that hypothesis **H1** is Entail.

Controlled Changes → Expected Model Response

Deleting the *row* **No. of Listing** should change the label for **H1** from Entail to Neutral.

3. Motivation

Any "evidence-based reasoning" system should demonstrate expected, predictable behavior in response to controlled changes to its inputs.

Case Study on Tabular Inference

4. Our Contributions

1. Systematic target probing can highlight the limitation of tabular reasoning models
2. Such targeted probes are data efficient and work with minimal to no supervision
3. Probing can highlight several problems in tabular reasoning models:

1. Use of annotation artifacts
2. Use of incorrect evidence
3. Overfitting on pre-trained knowledge

5. Systematic Probes

We define three types of systematic probes, as follows:

1. **Annotation Artifacts:** *Can a model make inference about a hypothesis without a premise?*

Yes, models largely rely on spurious correlation between hypothesis and inference label.

2. **Evidence Selection:** *Is the model drawing inferences based on right evidence in the premise?*

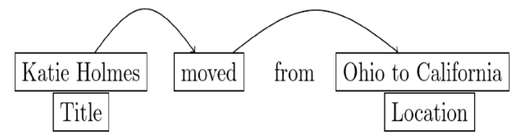
No, models do not look at correct evidence as required for right reasoning.

3. **Counterfactual Instances:** *How will the model react if the primary evidence is counterfactual to pre-trained data?*

Model relies on information from pre-trained language models rather than tabular evidence

6. Annotation Artifacts

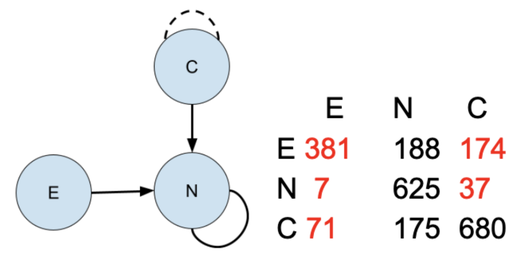
1. Modify the hypothesis in such a way that the inference label is retained or flipped
2. **Modified Expression Types:** named entity, numerical, temporal, quantification, lexical, negation, syntactic alternation, subjective.
3. E.g. for Named Entity modification
Katie Homes moved from **Ohio to California**



Katie Holmes moved from **South Africa To California**.
Expected Responce: Entail → Contradict

7. Evidence Selection

1. Alter the premise table via simple operations for deterministic change of inference label.
2. **Possible Operations:** row deletion, row-value updation, new row insertion, and row perturbation
3. E.g. for hypothesis **H1**
Delete the *row* **"No. of Listing"** from example table



Expected Responce (H1): Entail → Neutral

4. E.g. for hypothesis **H1**
Delete the *row* **"Location"** from example table
Expected Responce (H1): Entail → Entail

8. Counterfactual Instances

1. Update the premise table to include counterfactual data in order to retain or change the inference label.

Dressage	
Highest governing body	International Federation for Equestrian Sports (FEI)
Characteristics	
Contact	No
Team members	Individual and team at international levels
Mixed gender	Yes
Equipment	Horse, appropriate horse tack
Venue	Arena, indoor or outdoor
Presence	
Country or region	Worldwide
Olympic	1912
Paralympic	1996

Both men and women can complete in the **contactless** sport of Dressage → Entail

Dressage	
Highest governing body	International Federation for Equestrian Sports (FEI)
Characteristics	
Contact	Yes
Team members	Individual and team at international levels
Mixed gender	Yes
Equipment	Horse, appropriate horse tack
Venue	Arena, indoor or outdoor
Presence	
Country or region	Worldwide
Olympic	1912
Paralympic	1996

Both men and women can complete in the **contactless** sport of Dressage → Contradict

9. Inoculation Study

1. Can additional fine-tuning with the perturbed examples (i.e., data inoculation) help?
2. Model performance increases on challenge sets but degrades on the original α_1, α_2, and α_3 test sets.

No, changes in the data distribution during training have a negative impact on model performance.

11. Observation

1. **Artifacts:** Models rely on spurious correlation between hypothesis and inference label.
2. **Evidence:** Models does not look at correct evidence required for correct reasoning.
3. **Counterfactual:** Model relies on information from pre-trained language models rather than tabular evidence.
4. **Inoculation:** Changes in the data distribution during training have a ngeative impact on model.

Data and Software:
https://tabprobe.github.io