Incorporating External Knowledge to Enhance Tabular Reasoning

https://knowledge-infotabs.github.io/

J. Neeraja[1*], Vivek Gupta[2*], Vivek Srikumar[2]
[1]IIT Guwahati; [2]University of Utah

TABULAR INFERENCE PROBLEM

- The **tabular natural language inference** problem is similar to standard NLI
- But here, the **premises are tabular data**
- Task: to decide whether given hypothesis is **true** (entailment), **false** (contradiction) or **undetermined** (neutral) given a premise table

Check out InfoTabs (Gupta et al., 2020)
https://infotabs.github.io

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.

In this example from the InfoTabS dataset (Gupta et al., 2020),

H1: entailed ; **H2:** contradictory ; **H3:** neutral

MOTIVATION

Recent work for tabular reasoning focuses on **building** sophisticated **neural models**.

Questions?

- How models designed for the **raw text adapt for tabular data**?
- How to **represent data** and **incorporate knowledge** into model?
- Can better **preprocessing** of **tabular information** enhance model?

CHALLENGES

1. Poor Representation of Tabular Information
2. Missing Implicit Lexical Knowledge
3. Presence of Distracting Information
4. Missing Domain Knowledge about Keys

Can we fix these problems by changing how tabular information is provided to a standard RoBERTa model?

CHALLENGE: POOR TABLE REPRESENTATION

- In Gupta et al., 2020,
  Universal template: "The ***k*** of ***t*** are ***v***."

The Founded of New York Stock Exchange are May 17, 1792; 226 years ago.

- Most sentences are ungrammatical

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.

In this example from the InfoTabS dataset (Gupta et al., 2020),

H1: entailed ; **H2:** contradictory ; **H3:** neutral

SOLUTION: BETTER PARAGRAPH REPRESENTATION

- Entity specific templates : use value entity types DATE or CARDINAL or BOOL

New York Stock Exchange was founded on May 17, 1792; 226 years ago.

- Add category information

New York Stock Exchange is an **organization**.

More grammatical and meaningful sentences

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.

In this example from the InfoTabS dataset (Gupta et al., 2020),

H1: entailed ; **H2:** contradictory ; **H3:** neutral

CHALLENGE: MISSING IMPLICIT LEXICAL KNOWLEDGE

- Limited training data
- Interpreting hypernym words like *'fewer'*, and *'over'* and negations like '*never'* or *'not'*.

e.g.
H2`: Fewer than 2,500 stocks are listed in the NYSE
H2`: contradictory

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.

In this example from the InfoTabS dataset (Gupta et al., 2020),

H1: entailed ; **H2:** contradictory ; **H3:** neutral

SOLUTION: IMPLICIT KNOWLEDGE ADDITION

Can pre-training on large dataset help?

- Pre-training with MNLI data
- Then, fine-tune on InfoTabS

Exposes model to diverse lexical constructions
Representation is better tuned for the NLI task

e.g.
H2`: Fewer than 2,500 stocks are listed in NYSE
H2`: entailed

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.

In this example from the InfoTabS dataset (Gupta et al., 2020),

H1: entailed ; **H2:** contradictory ; **H3:** neutral

CHALLENGE: PRESENCE OF DISTRACTING INFORMATION

- Given hypothesis, limited rows are relevant
  - In **H1** and **H2**, row with key ***No. of listings*** is sufficient.
  - Similarly, for **H3**, row with key ***Volume*** is sufficient.
- BERT also has tokenization limit
  - longer tables are cropped

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.

In this example from the InfoTabS dataset (Gupta et al., 2020),

H1: entailed ; **H2:** contradictory ; **H3:** neutral

SOLUTION: DISTRACTING ROW REMOVAL

Select only rows relevant to hypothesis

Use alignment based retrieval algorithm with fastText vectors (Yadav et al. (2019, 2020))

E.g. for H1 & H2, prune table to

New York Stock Exchange	
No. of listings	2,400

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.

In this example from the InfoTabS dataset (Gupta et al., 2020),

H1: entailed ; **H2:** contradictory ; **H3:** neutral

CHALLENGE: MISSING DOMAIN KNOWLEDGE ABOUT KEYS

For **H3**, we need to interpret the key ***Volume*** in the **financial context.**

✅ In **capital markets**, ***volume*** is the total number of a security that was traded during a given period of time.

rather than

❌ In **thermodynamics**, the **volume** of a system is an extensive parameter for describing its thermodynamic state.

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.

In this example from the InfoTabS dataset (Gupta et al., 2020),

H1: entailed ; **H2:** contradictory ; **H3:** neutral

SOLUTION: EXPLICIT KNOWLEDGE ADDITION

Add ***explicit*** information to ***enrich keys***

This improves model's ability to disambiguate meaning of keys

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.

In this example from the InfoTabS dataset (Gupta et al., 2020),

H1: entailed ; **H2:** contradictory ; **H3:** neutral

SOLUTION: EXPLICIT KNOWLEDGE ADDITION

Approach

- Use **BERT** on wordnet examples to find key embeddings
- Get key embeddings from premise using **BERT**
- Find the **best match** and add it definition.

For H3, add to the table in the end:

Volume: total number of a security that was traded during a given period of time.

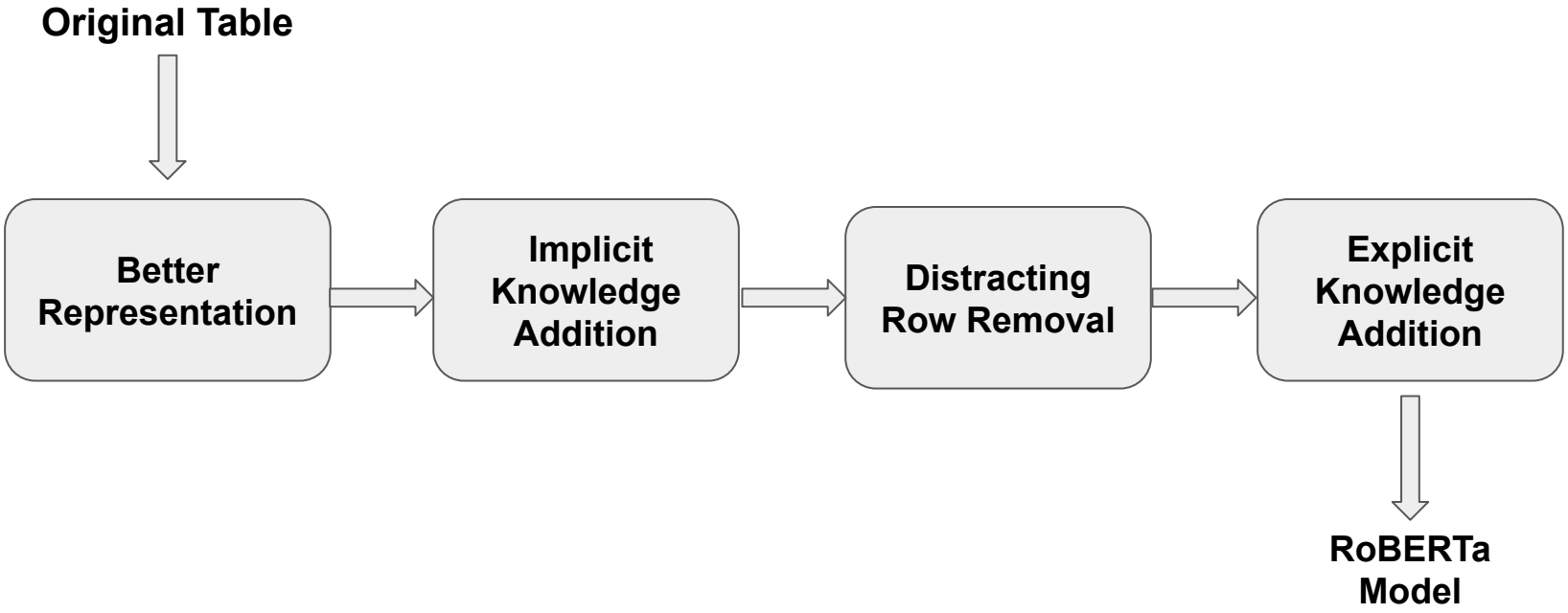
New York Stock Exchange	
Type	Stock exchange
Location	New York City, New York, U.S.
Founded	May 17, 1792; 226 years ago
Currency	United States dollar
No. of listings	2,400
Volume	US$20.161 trillion (2011)

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.

In this example from the InfoTabS dataset (Gupta et al., 2020),

H1: entailed ; **H2:** contradictory ; **H3:** neutral

PROPOSED SOLUTION



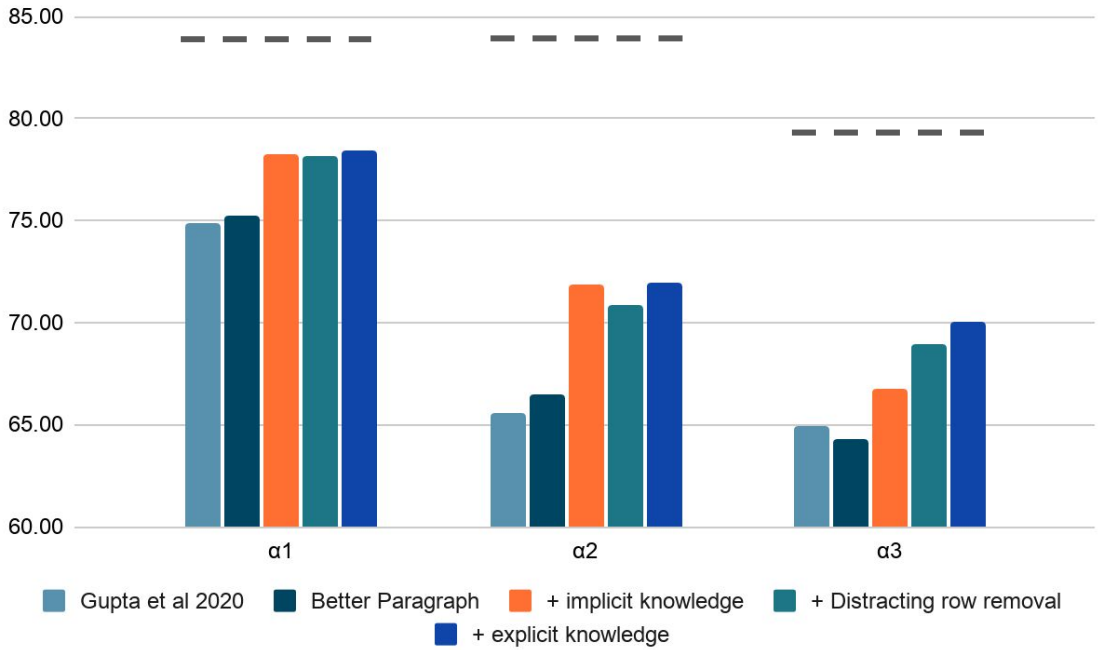
RESULTS AND ANALYSIS

InfoTabS dataset splits :

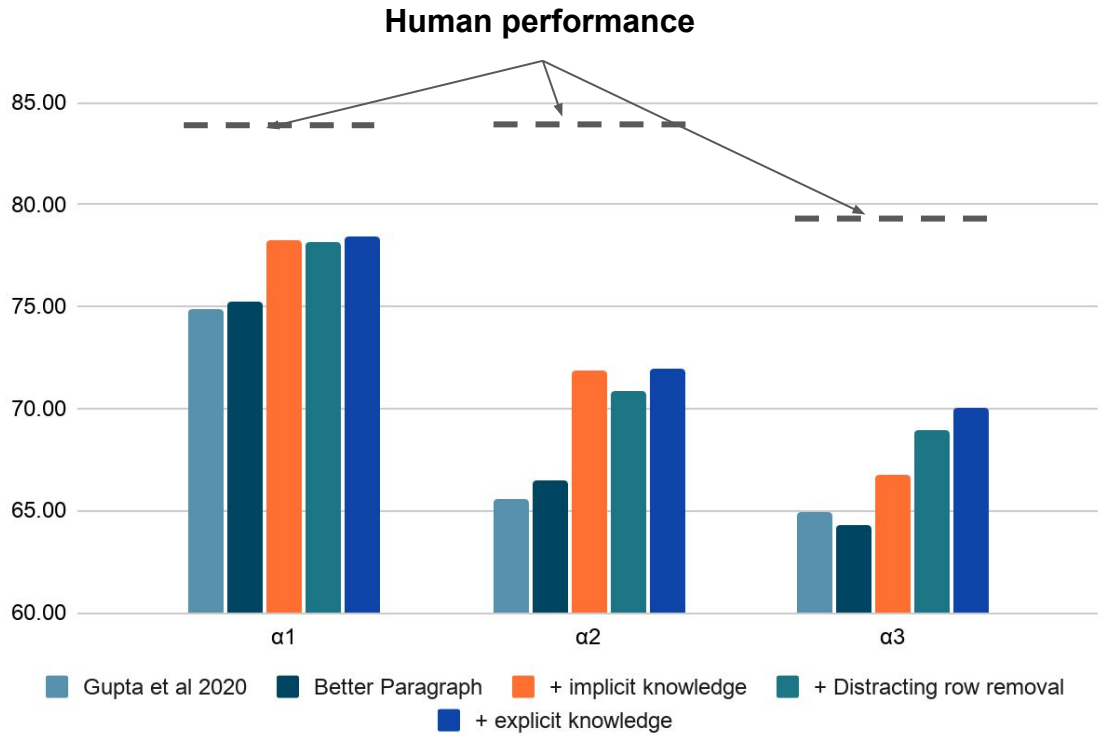
- **α1** contains table from same domain (similar to dev & train set)
- **α2** has examples from same domain but entail-contradict label (e.g. ‘over’ to ‘under’) flipped by minimal change i.e. **adversarial**.
- **α3** is **zero-shot** cross domain tables (exclusive from train set domains)

Check out InfoTabS: https://infotabs.github.io

RESULTS AND ANALYSIS



RESULTS AND ANALYSIS

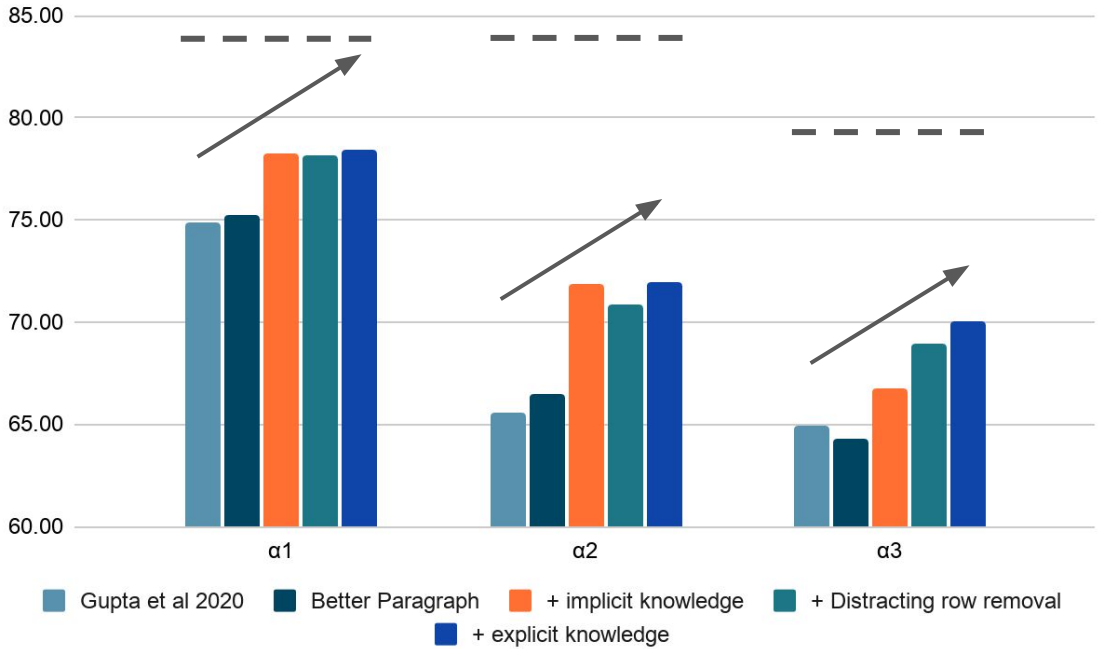


RESULTS AND ANALYSIS

Observation

Overall **Pre-processing** improves performance

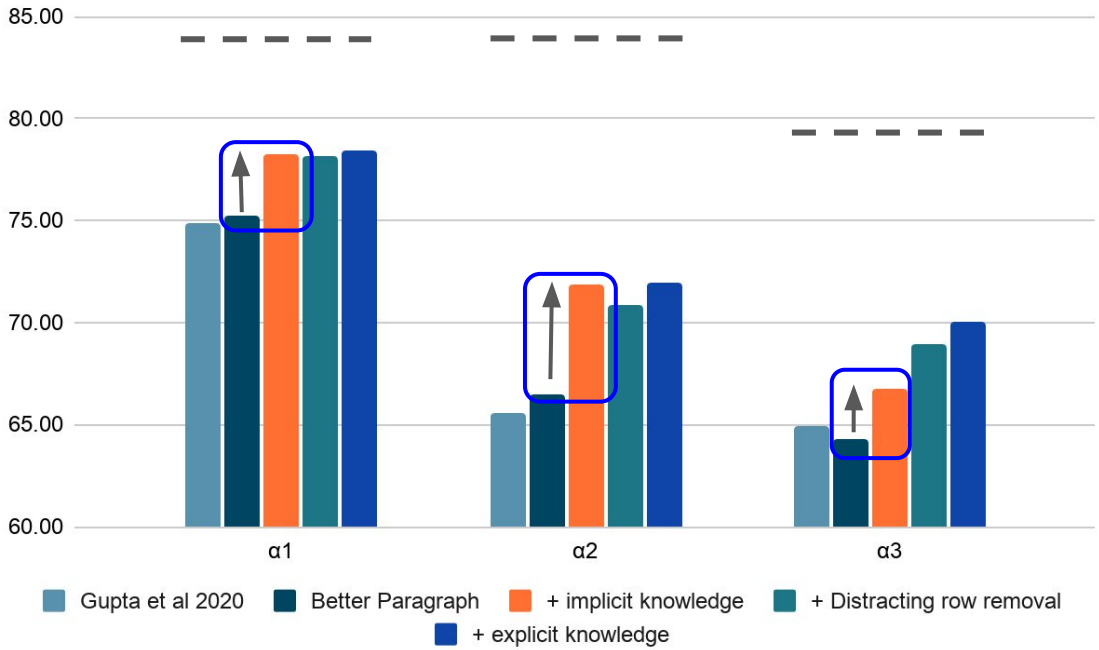
Ablation : all changes needed, knowledge is the most important



RESULTS AND ANALYSIS

Observations

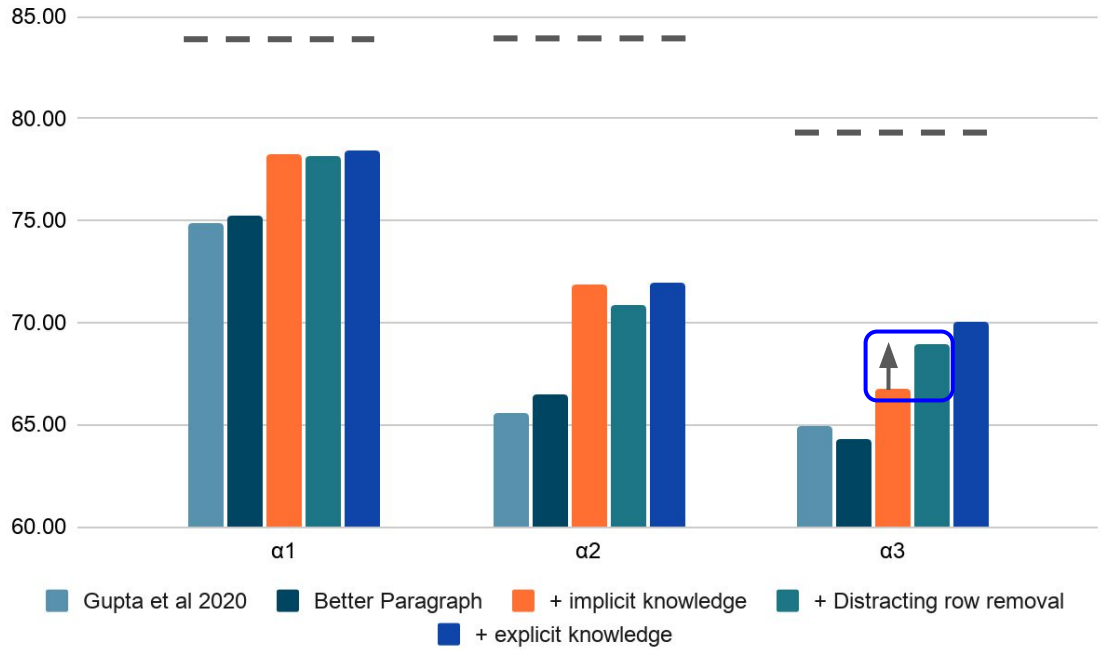
- **Implicit knowledge** improve **α1, α2** & **α3**
- reflect model learning implicit knowledge



RESULTS AND ANALYSIS

Observations

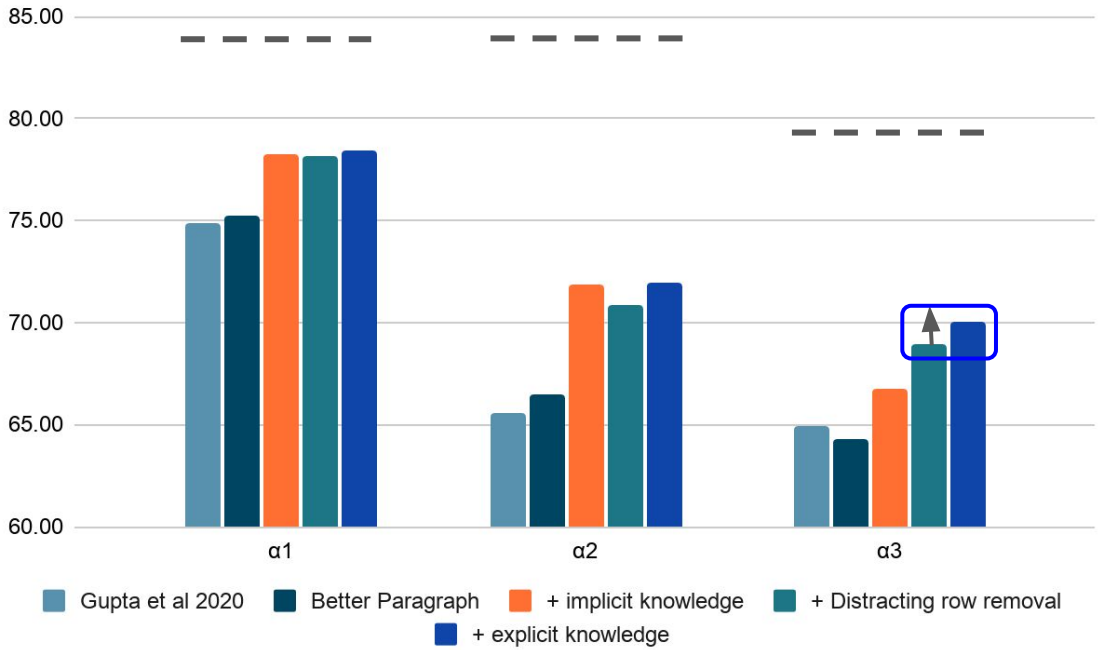
- **Distracting row removal** improve **α3**
- α3 longer tables → BERT tokenization prune relevant rows



RESULTS AND ANALYSIS

Observations

- **+explicit** knowledge help α3 improvement
- α3 is zero-shot → keys information is needed.



CONCLUSION

- Effective preprocessing techniques **improve tabular reasoning**, case study on **tabular inference show performance improvement**.
- Proposed preprocessing lead to **significant improvements** especially on the **adversarial** sets of tabular inference dataset (InfoTabS).
- Solutions applicable to *question answering* and *generation* problems which involve both *tabular and textual inputs*, especially for **adversarial** evaluation.
- We recommend that modifications should be **standardized across other table reasoning tasks**.

Check out Knowledge_InfoTabs: https://knowledge-infotabs.github.io/