Information Extraction for Trustworthy Tabular Inference

Vivek Gupta[1], Shuo Zhang[2], Alakananda Vempala[2], Yujie He[2], Temma Choji[2], Vivek Srikumar[1]

[1]University of Utah; [2]Bloomberg

1. Tabular Inference

- Inference task where premises are tabular in nature
- Given a premise table determine if hypothesis is true (*entailment*), false (*contradiction*), or undetermined (*neutral*), i.e., tabular natural language inference.
- Below is an example premise table from InfoTabS dataset (Gupta et al., 2020). Here, the hypothesis **H1**: entails ; **H2**: contradicts ; **H3**: 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.

2. The Problem

- Model does not provide the **inference evidence** and the **reasoning steps**.
- From the example above, the *row* **No. of Listing** is required to establish that hypothesis **H1** and **H2** are *entail* and *contradict* respectively.
- **Deletion Probing** (Gupta et al., 2022) shows that deleting the *row* **no. of listing** change **H1** and **H2** predictions to *neutral*.

3. Motivation

- *Not enough* for a model to be merely *right*, but also *right for the right reasons*.
- *Identifying* the *relevant elements of input* as the *right reasons* is essential for correct *tabular reasoning*.

4. Our Contributions

- Tabular reasoning models ignore the evidence table rows (Gupta et al., 2022)
- Model should not be just right, but right for the right reasons (this work)

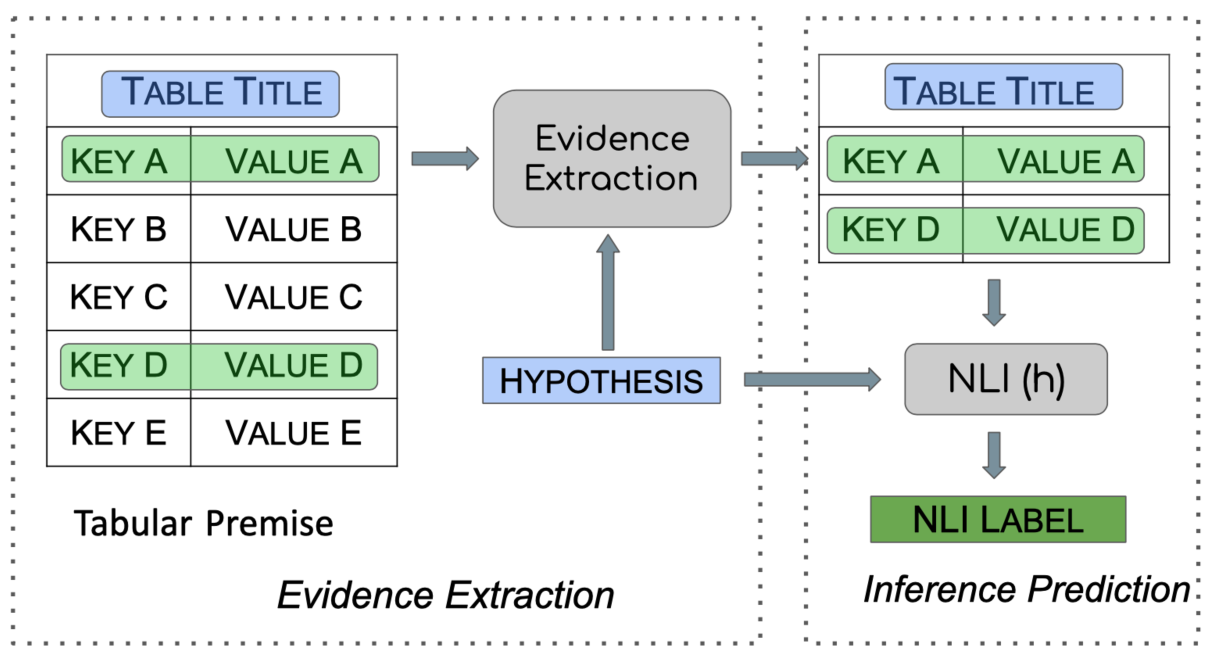
Case Study on InfoTabS

✗ Tabular Reasoning → ✓ Trustworthy Tabular Reasoning
✗ Only Inference → ✓ Evidence Extraction + Inference

- Two-fold benefits
  1. make model trustworthy
  2. benefit the reasoning task

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

5. Trustworthy Tabular Inference



Two-stage Approach

Modified Model Predictions for H1

- inference label: *entail* label (as earlier)
- evidence (i.e., the relevant rows):
  No. of Listings : 2,400

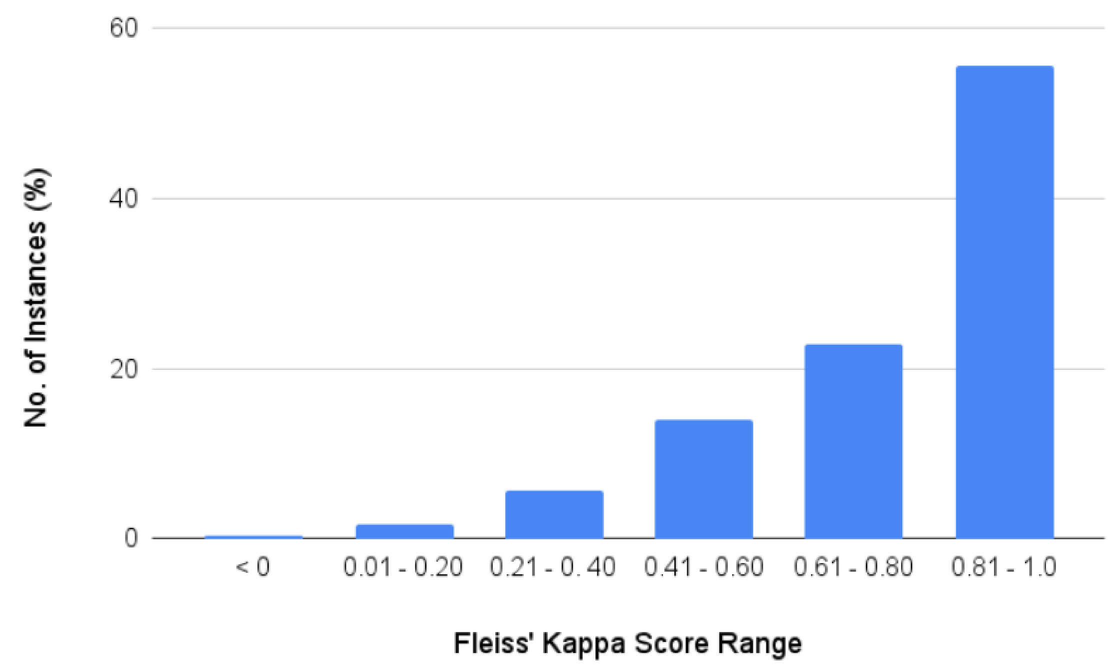
6. Evidence Extraction

- Unsupervised methods: Textual similarity
  - DRR (Neeraja et al., 2021)
  - WMD (Gupta et al., 2020)
  - SimAlign (Sabet et al., 2020)
  - SimCSE (Gao et al., 2021)
- Supervised methods: Binary classification

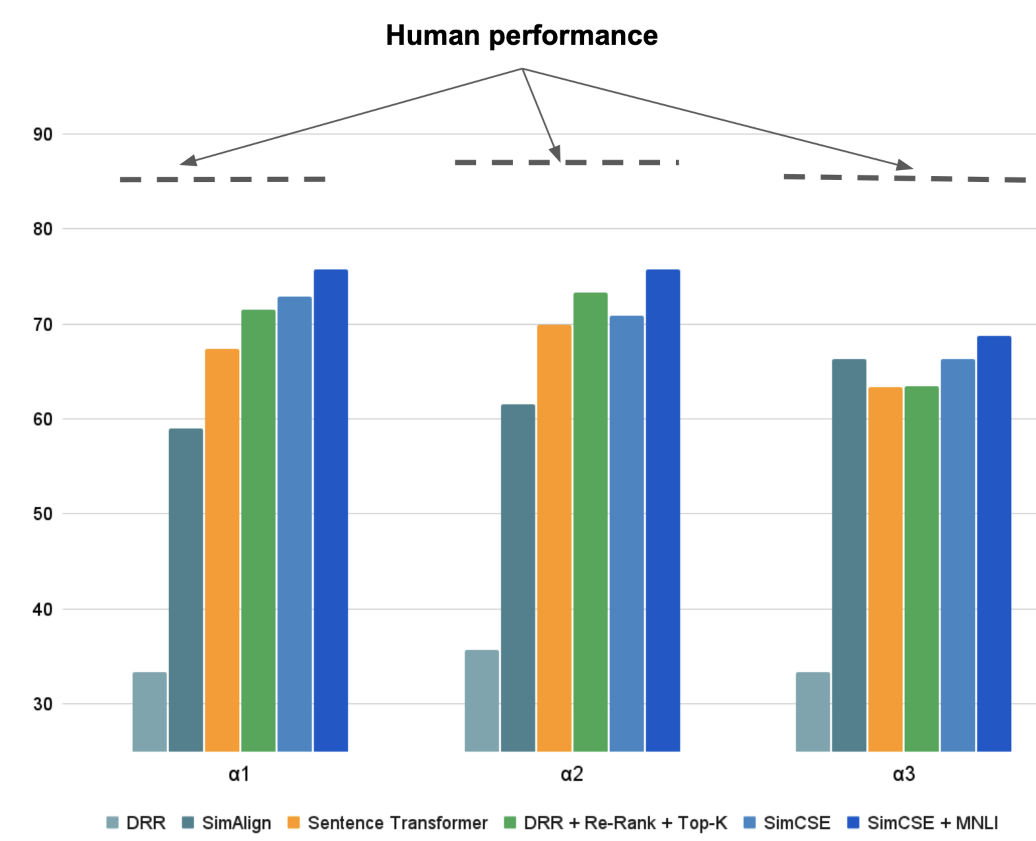
f(row, hypothesis) → {relevant, irrelevant}
Hard negative via unsupervised models

7. Human Annotation Agreement

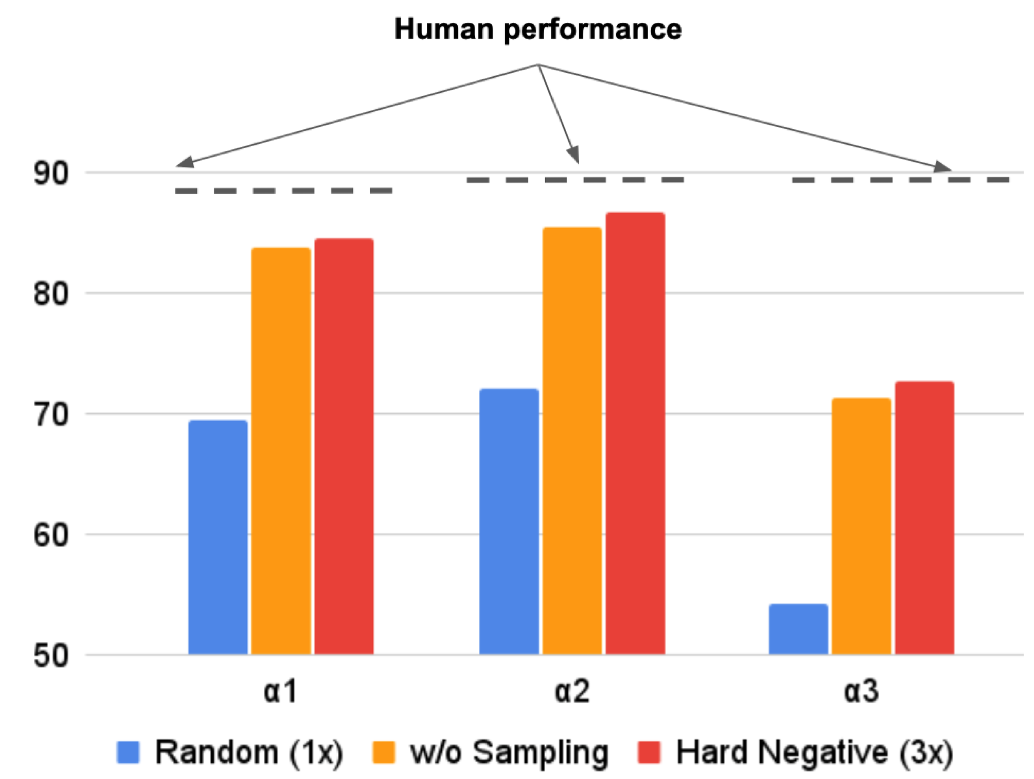
F1-Score	Precision	Recall	F1-Score
Avg-Micro	89.28	91.57	90.06
Avg-Macro	89.24	89.75	89.49



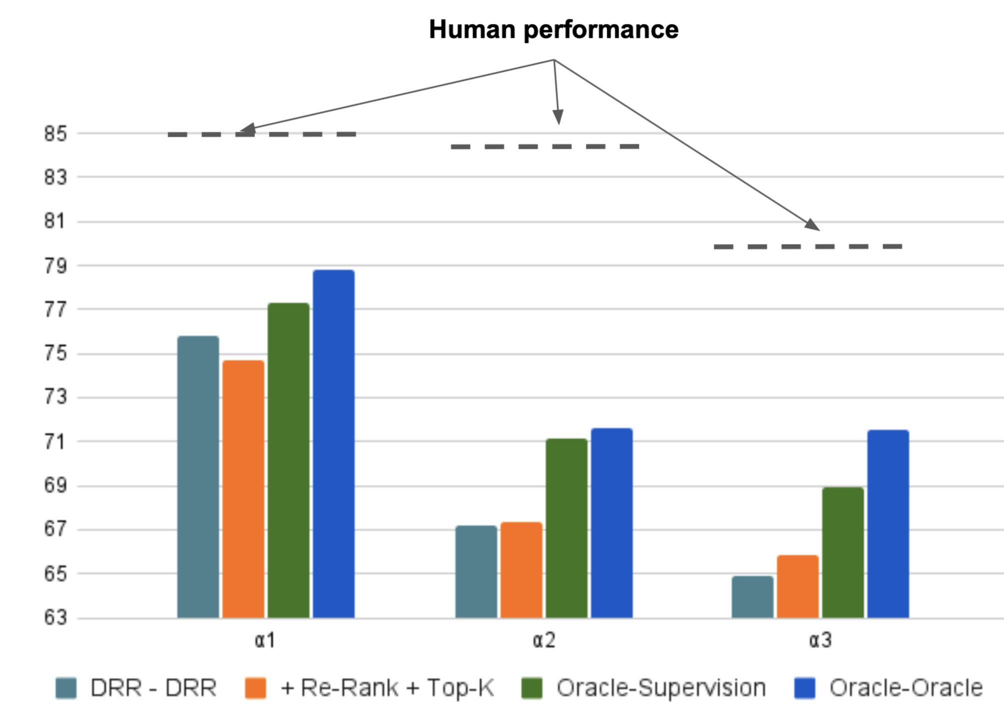
8. Unsupervised Extraction



9. Supervised Extraction



10. Final Inference



11. Observations

1. Unsupervised extraction **re-rank** and **dynamic top-K** benefits
2. Unsupervised extraction **Hypo-title swap** confounding of <TITLE> similarity beneficial.
3. Supervised extraction **significant better** than unsupervised extraction
4. Adding **hard negative (3x)** better than **random (1)** or **no sampling**.
5. Beneficial for NLI especially on **zero-shot (out-of-domain)** α_3 dataset