Table Unionability Is Uncertain and That's Why Humans and AI Need Each Other

Based on: *Nina Klimenkova*, *Sreeram Marimuthu, Roee Shraga.*
"Humans, Machine Learning, and Language Models in Union: A Cognitive Study on Table Unionability". HILDA at SIGMOD 2025

1. Motivation & Background

Table Unionability: a fundamental challenge in data discovery - identifying tables that can be meaningfully combined [3].

Evolving Definitions:
- Traditional
- Relaxed [1]
- Relation-based [2]
- Context-aware [3]

Cognitive Challenge:
- Semantic & Context interpretation
- Domain knowledge
- Judgment under ambiguity

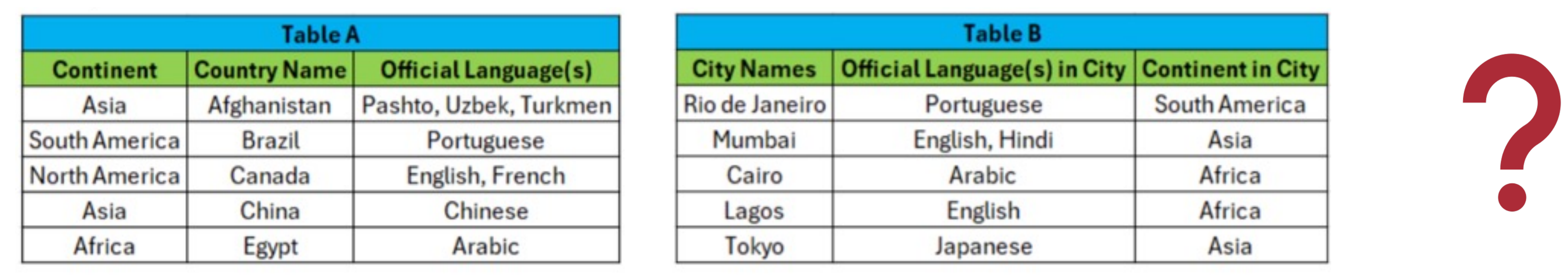
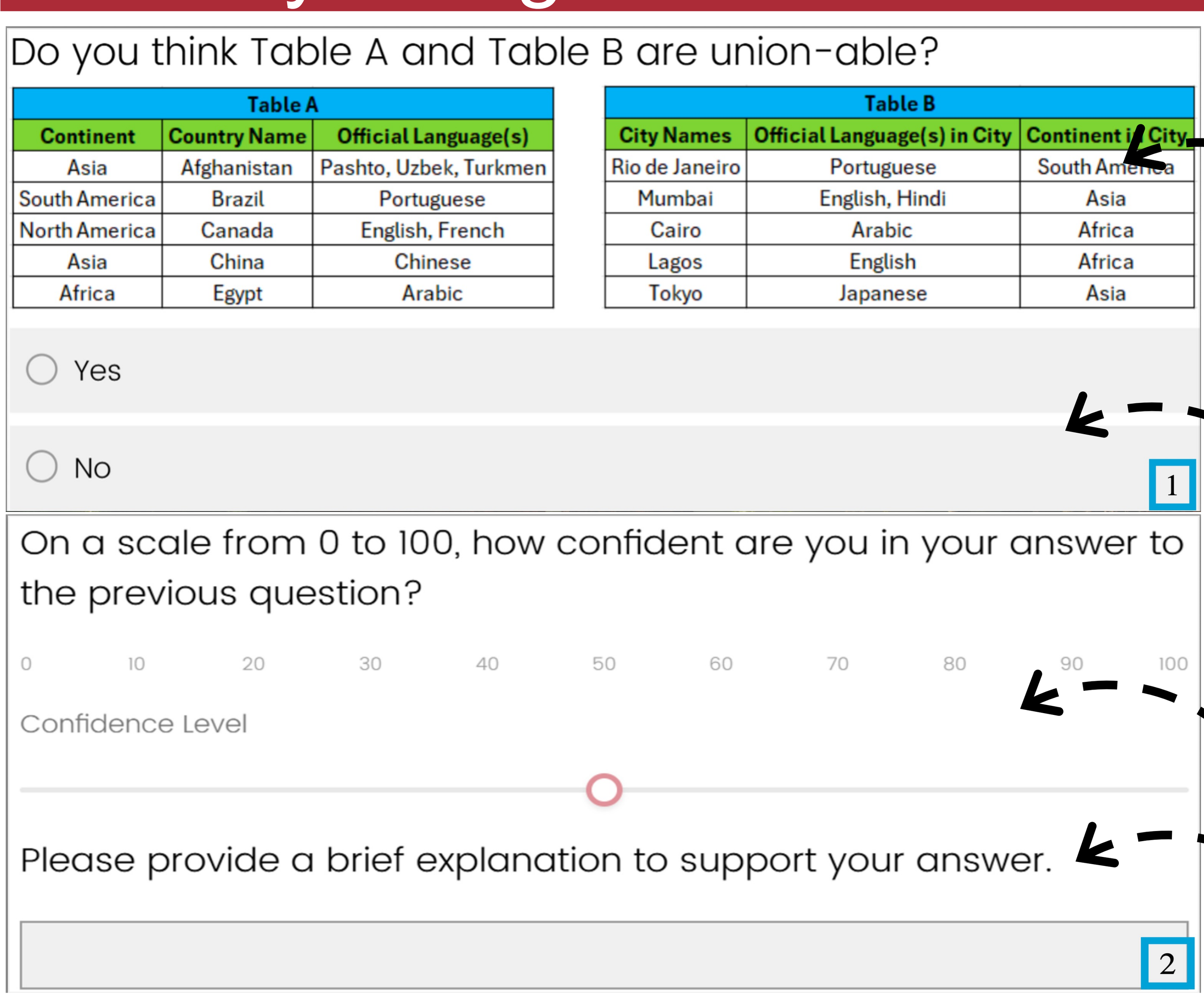


Table A		
Continent	Country Name	Official Language(s)
Asia	Afghanistan	Pashto, Uzbek, Turkmen
South America	Brazil	Portuguese
North America	Canada	English, French
Asia	China	Chinese
Africa	Egypt	Arabic

Table B		
City Names	Official Language(s) in City	Continent in City
Rio de Janeiro	Portuguese	South America
Mumbai	English, Hindi	Asia
Cairo	Arabic	Africa
Lagos	English	Africa
Tokyo	Japanese	Asia

Can we use human input patterns to improve the quality of table unionability judgments?

2. Study Design and Behavioral Observations



Do you think Table A and Table B are union-able?

Yes

No

On a scale from 0 to 100, how confident are you in your answer to the previous question?

0 10 20 30 40 50 60 70 80 90 100

Confidence Level

Please provide a brief explanation to support your answer.

4 survey versions

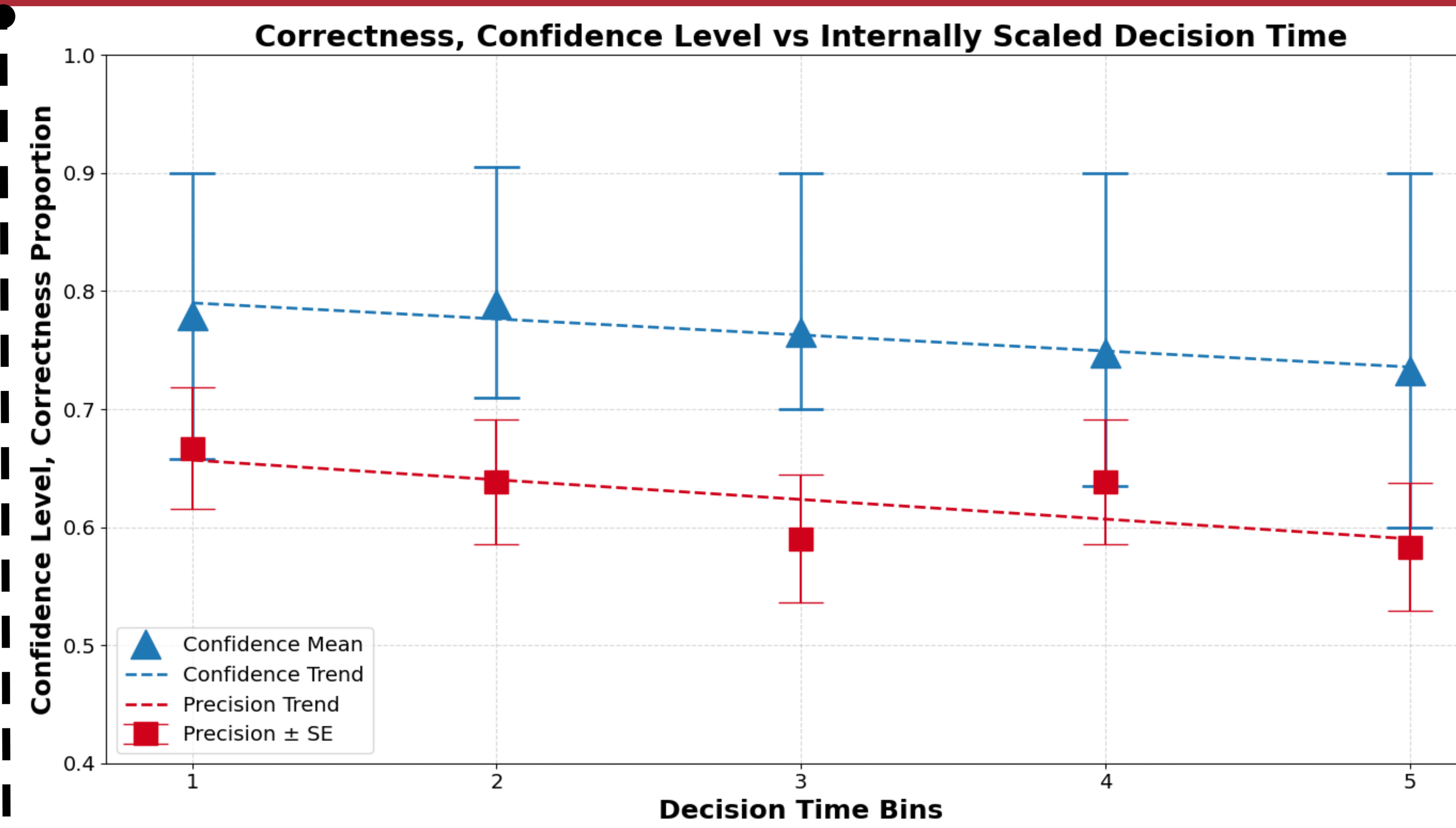
Tables from UGEN dataset [4]

Unionability judgment

Confidence

Explanation by user

Behavioural tracking: clicks, decision time, interaction patterns



- Confidence decreases with decision time (0.79 → 0.74)
- Accuracy drops with longer decisions (0.66 → 0.59)

→ Suggests overthinking may hurt performance

→ Longer deliberation = harder cases

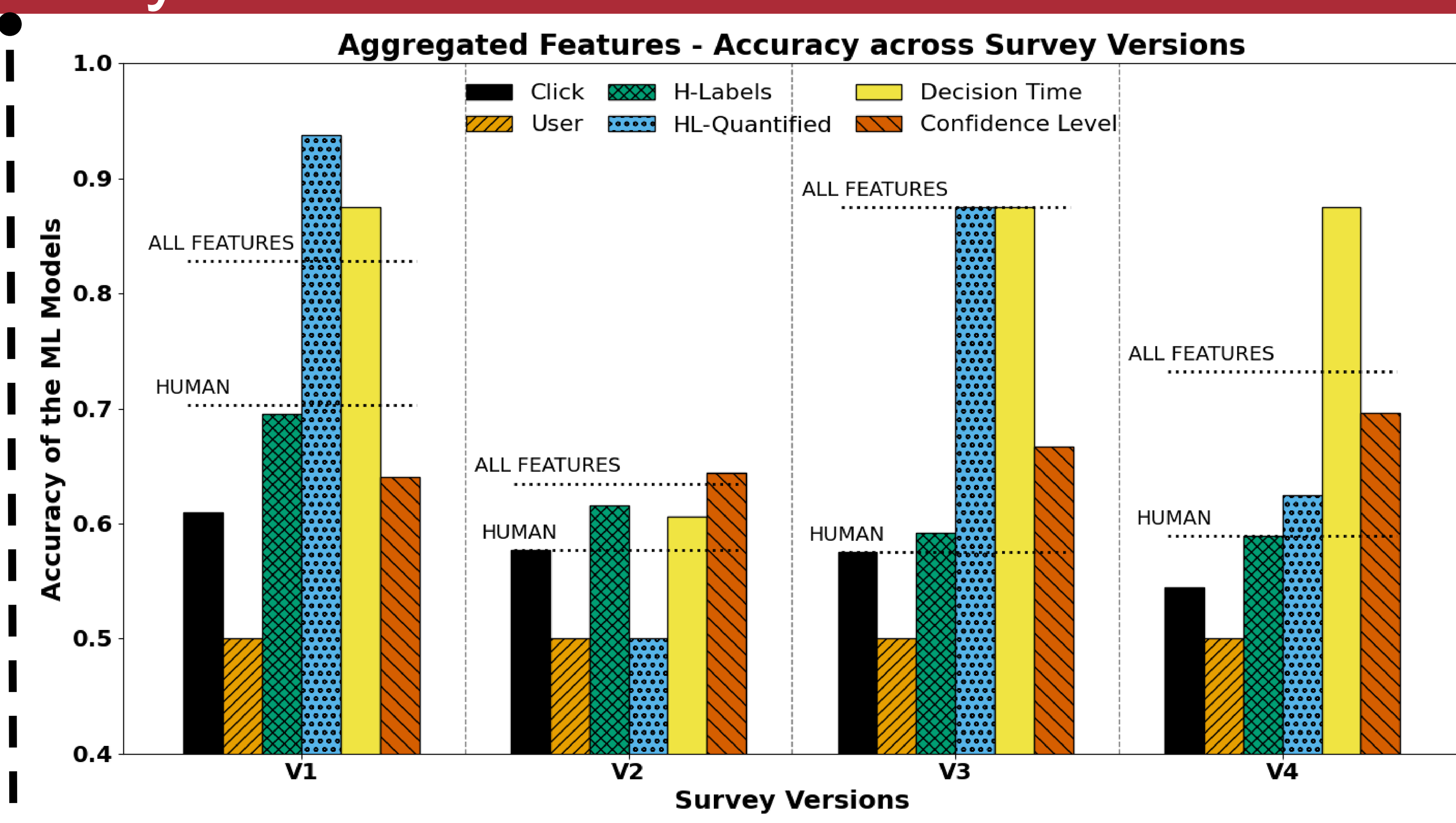
3. Approach: Calibrating Human Table Unionability Labels

11 Initial features → 33 Engineered features → 6 Aggregated subsets → Train classifiers (LR, KNN, RF, XGB) → Calibrated labels

Feature Group	Description
Click	Click behavior metrics
User	Demographics & metadata
Human-Labels	Participant response items
Quantified-Human-Labels	Group-level correctness
Decision-Time	Temporal decision measures
Confidence Level	Self-reported confidence

Goal: predict if a human's answer is correct → cleaner labels

Metric: accuracy (Yes = 1, No = 0)



+32.3% Decision Time Features

+11.6% Confidence Level

+20.1% Quantified Labels

-18.1% User Demographics only

4. Experiments and Results

Survey	Human		ML (All Features)	LLM	Human+LLM
	Acc.	Majority	Acc.	Actual	Added Ctx.
V1	0.70	**1.00 (+42.2%)**	0.83 (+17.8%)	0.63 (-11.1%)	0.75 (+6.7%)
V2	0.58	0.50 (-13.3%)	0.64 (+10.1%)	0.50 (-13.4%)	**0.63 (+8.3%)**
V3	0.58	**0.88 (+52.2%)**	0.88 (+52.2%)	0.63 (+8.7%)	**0.88 (+52.2%)**
V4	0.59	0.63 (+6.1%)	0.73 (+24.2%)	0.63 (+6.1%)	**0.75 (+27.3%)**
Avg.	0.61	**0.75 (+22.7%)**	0.77 (+25.5%)	0.59 (-2.8%)	**0.75 (+22.7%)**

- Raw human judgments are inconsistent
- Behavior-aware ML and collective human input improve accuracy
- LLMs alone do not consistently outperform humans

References

[1] Fatemeh Nargesian, Erkang Zhu, Ken Q. Pu, and Renée J. Miller. Table union search on open data. VLDB 2018

[2] Aamod Khatiwada, Grace Fan, Roee Shraga, Zixuan Chen, Wolfgang Gatterbauer, Renée J. Miller, and Mirek Riedewald. Santos: Relationship-based semantic table union search. SIGMOD 2023

[3] Fan, Grace, Jin Wang, Yuliang Li, Dan Zhang, and Renée J. Miller. "Semantics-Aware Dataset Discovery from Data Lakes with Contextualized Column-Based Representation Learning." VLDB 2023

[4] Koyena Pal, Aamod Khatiwada, Roee Shraga, and Renée Miller. 2023. Generative Benchmark Creation for Table Union Search.

Nina Klimenkova | PhD Student

Department of Data Science, Worcester Polytechnic Institute