Understanding Human Judgment in Table Unionability

NEDB Day 2026 • Boston • January 16, 2026

Sreeram Marimuthu, Nina Klimenkova, Roee Shraga

WPI

Background

Table Unionability: a fundamental challenge in data discovery - identifying tables that can be meaningfully combined (unioned)

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

Query table

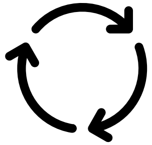
Table B		
City Names	**Official Language(s) in City**	**Continent in City**
Rio de Janeiro	Portuguese	South America
Mumbai	English, Hindi	Asia
Cairo	Arabic	Africa
Lagos	English	Africa
Tokyo	Japanese	Asia

Datalake table

*In this work, we focus on **judging unionabilty** rather than the search itself*

Background

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



Evolving Definitions

- Traditional: **All** columns should be unionable
- Relaxed [1]: **Some** columns should be unionable
- Relationship-based [2]: **Some (meaningful)** columns should be unionable
- Context-aware [3]: **Some (context-consistent)** columns should be unionable

[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

Motivation

Evolving Definitions

- Traditional
- Relaxed [1]
- Relationship-based [2]

Cognitive Challenge

- Semantic interpretation
- Context understanding
- Domain knowledge
- Judgment under ambiguity

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

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

[1] Fatemeh Nargesian, Erkang Zhu, Ken Q. Pu, and Renée J. Miller. Table union search on open data. Proceedings of the VLDB Endowment (PVLDB), 11(7):813– 825, 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. Proceedings of the ACM onManagement of Data, 1(1):1–25, 2023

Survey Design

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

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

◯ Yes

◯ No

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

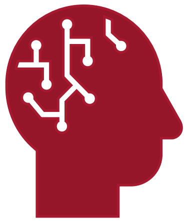
0 10 20 30 40 50 60 70 80 90 100

Confidence Level

Please provide a brief explanation to support your answer.

2

- ✓ 4 survey versions for balanced design



- ✓ Behavioral tracking: clicks, decision time, interaction patterns
- ✓ Tables from UGEN benchmark dataset [4]

[4] Koyena Pal, Aamod Khatiwada, Roee Shraga, and Renée J. Miller. Alt-gen: Benchmarking table union search using large language models. In Proceedings of the VLDB 2024 Workshop: Tabular Data Analysis Workshop (TaDA), 2024. Available at: https://github.com/northeastern-datalab/gen.

Participant Demographics

58
Participants

8
Questions

464
Total Responses

- Students in CS, Data Science, AI
- Undergraduate, Masters, PhD levels
- 81% majoring in data/computing fields
- 70%+ fluent/native English speakers

Dataset

11 Initial features

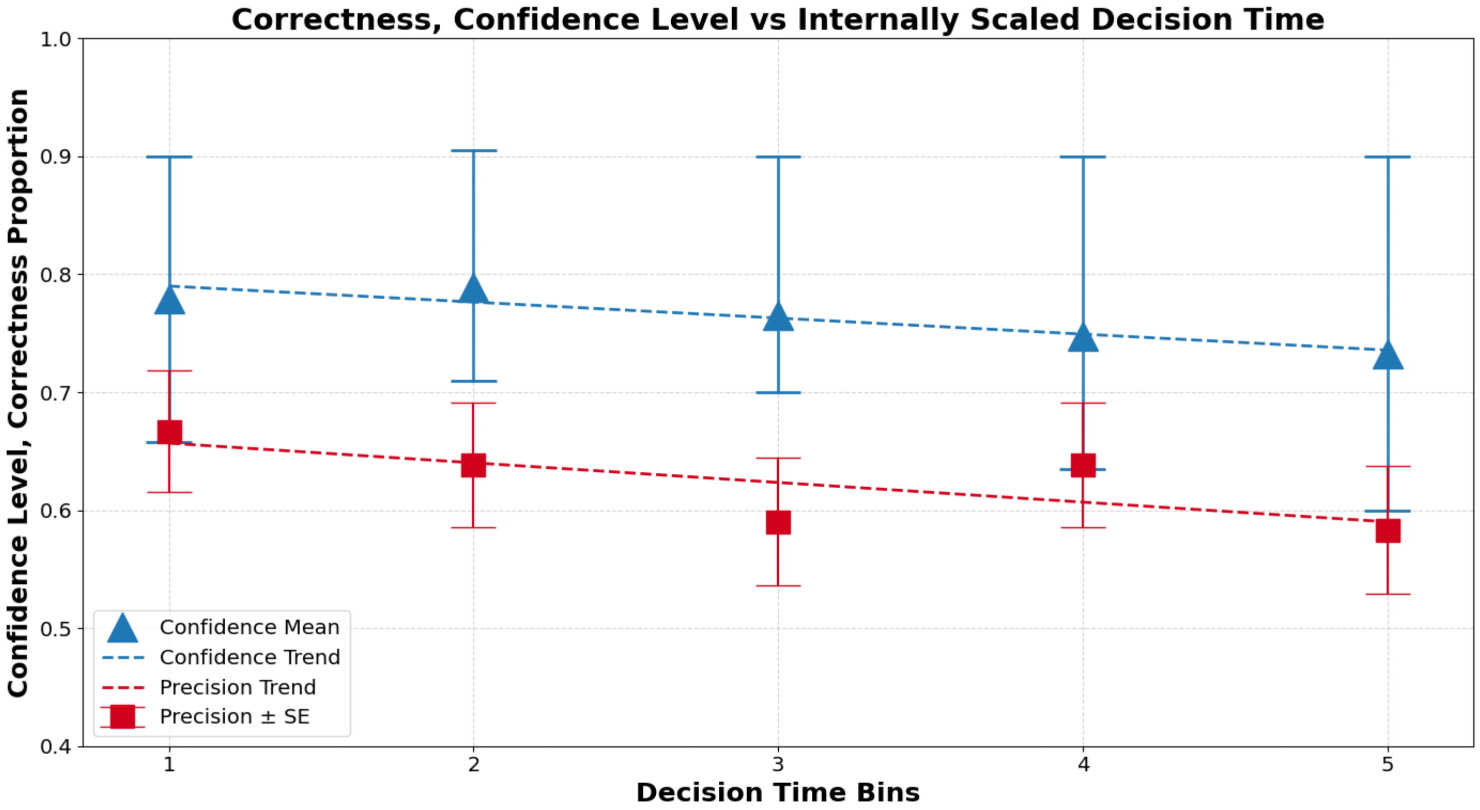
33 Engineered features

6 Aggregated subsets

6 Aggregated subsets

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

Human Behavior Analysis



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

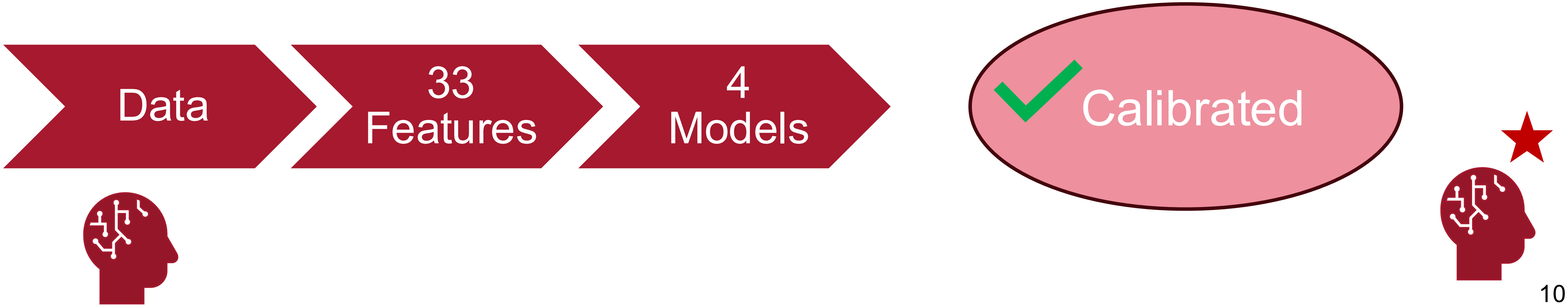
→ Suggests overthinking may hurt performance

→ Longer deliberation = harder cases

Calibrating Human Table Unionability Labels

Approach

- Train 4 classifiers: LR, KNN, RF, XGB
- 33 features (3 versions) → test on held-out version
- **Goal: predict whether a human answer is correct**
  → cleaner labels
- Metric: accuracy (Yes=1, No=0)



Calibrating Human Table Unionability Labels

Version	Human Baseline	ML Enhanced	Improvement	Best Model
V1	0.70	**0.83**	+17.8%	Logistic Regression
V2	0.58	**0.64**	+10.1%	K-Nearest Neighbors
V3	0.58	**0.88**	+52.2%	Random Forest
V4	0.59	**0.73**	+24.2%	XGBoost
Average	0.61	**0.77**	+25.5%	-

Feature Group Performance

Average Improvement over Human Baseline:

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

+32.3% Decision Time Features

+11.6% Confidence Level

+20.1% Quantified Labels

-18.1% User Demographics only

Human-AI Collaboration

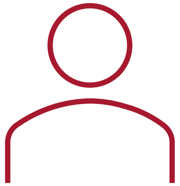
Tested Llama-3.3 70B with varying levels of human context

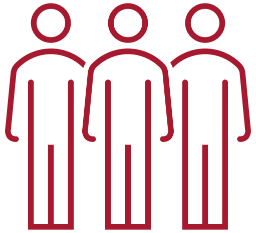
Scenario 1. Human (Actual): raw human responses

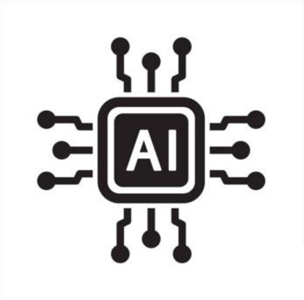
Scenario 2. Human (Majority): majority vote

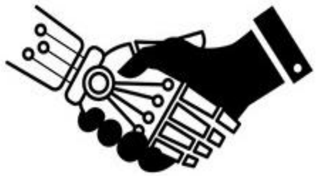
Scenario 3. LLM Only: just table description

Scenario 4. LLM + Human Context: added metacognitive data



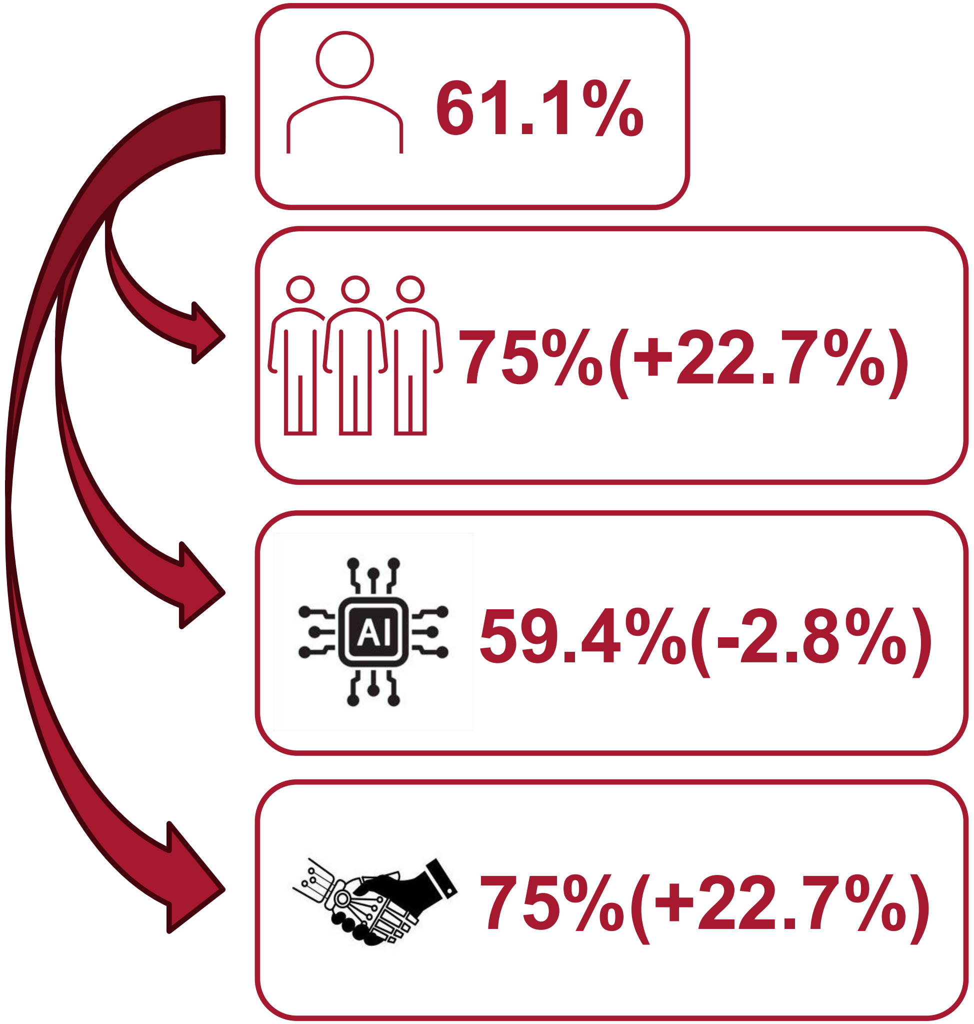






Human-AI Collaboration

1. Human (Actual)
2. Human (Majority)
3. LLM Only
4. LLM + Human Context



Key Takeaways

✓ Humans show systematic patterns in unionability decisions

✓ Behavioral features can improve label quality by 25%+

✓ LLMs benefit significantly from human context, but did not consistently improve through the addition of meta-cognitive factors

✓ Collective intelligence outperforms individual judgments

Thank you for listening!

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

rshraga@wpi.edu
nklimenkova@wpi.edu

YOSSI Lab