

# Human-Centered Exploration of Table Unionability [Experiment, Analysis & Benchmark]

Nina Klimenkova\*  
Worcester Polytechnic Institute  
Worcester, Massachusetts, USA  
nklimenkova@wpi.edu

Sreeram Marimuthu\*  
Worcester Polytechnic Institute  
Worcester, Massachusetts, USA  
smarimuthu@wpi.edu

Roe Shraga  
Worcester Polytechnic Institute  
Worcester, Massachusetts, USA  
rshruga@wpi.edu

## ABSTRACT

Table union search (TUS) identifies tables that can be meaningfully combined with a given query table and is a core task in data discovery over data lakes. Yet, what it means for two tables to be “unionable” is inherently ambiguous: domain experts may disagree even on seemingly simple cases, and existing benchmarks collapse this disagreement into binary labels, omitting the behavioral context behind human decisions. We take a human-centered view of table unionability and study how humans, traditional TUS methods, and large language models (LLMs) interact on this task. We introduce TUNE (Table UNIONability with human Evaluation), a benchmark of 464 expert judgments over 26 table pairs that records binary decisions, confidence scores, decision times, interaction traces, textual explanations, and post-survey reflections. Using TUNE, we (i) characterize human performance, overconfidence, and metacognitive quality (calibration and resolution); (ii) benchmark state-of-the-art TUS methods (Starmie, SANTOS, D3L), revealing complementary strengths and systematic misalignment with human judgments; and (iii) evaluate four experimental scenarios that combine human behavioral signals and TUS features using classical ML models and LLMs. The best configuration we experimented with reaches 84% accuracy, improving over both human majority vote and the strong standalone TUS method, while LLMs act as useful second opinions but are sensitive to conflicting signals. Overall, our results suggest that unionability labels reflect a structured yet imperfect human decision process and that hybrid human-model (may it be traditional classifiers or LLMs) pipelines provide more reliable and interpretable unionability assessments.

## PVLDB Reference Format:

Nina Klimenkova, Sreeram Marimuthu, and Roe Shraga. Human-Centered Exploration of Table Unionability [Experiment, Analysis & Benchmark]. PVLDB, 14(1): XXX-XXX, 2020.  
doi:XX.XX/XXX.XX

## PVLDB Artifact Availability:

The source code, data, and/or other artifacts have been made available at [https://github.com/NinaKlimenkova/TUNE\\_Benchmark](https://github.com/NinaKlimenkova/TUNE_Benchmark).

\*Both authors contributed equally to this research.

This work is licensed under the Creative Commons BY-NC-ND 4.0 International License. Visit <https://creativecommons.org/licenses/by-nc-nd/4.0/> to view a copy of this license. For any use beyond those covered by this license, obtain permission by emailing [info@vldb.org](mailto:info@vldb.org). Copyright is held by the owner/author(s). Publication rights licensed to the VLDB Endowment.

Proceedings of the VLDB Endowment, Vol. 14, No. 1 ISSN 2150-8097.  
doi:XX.XX/XXX.XX

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

Table A			Table B		
Continent	Country Name	Official Language(s)	City Names	Official Language(s) in City	Continent in City
Asia	Afghanistan	Pashto, Uzbek, Turkmen	Rio de Janeiro	Portuguese	South America
South America	Brazil	Portuguese	Mumbai	English, Hindi	Asia
North America	Canada	English, French	Cairo	Arabic	Africa
Asia	China	Chinese	Lagos	English	Africa
Africa	Egypt	Arabic	Tokyo	Japanese	Asia

☐ Yes

☐ No

1

Figure 1: Ambiguous unionability case from the survey.

## 1 INTRODUCTION

Most analytics tasks do not end with one table. Data scientists routinely augment their datasets by discovering and combining relevant external sources. This process is often framed as *data discovery* [1, 6, 8, 13, 19], a broad retrieval task focused on identifying potentially useful datasets in large repositories or data lakes. Within this broader landscape, TUS is a specialized discovery task that identifies candidate tables that can be meaningfully unioned with a given query table to expand its coverage or completeness [2, 9, 14, 16, 18, 20, 29, 32]. In this paper, we focus specifically on TUS and the challenges involved in defining, evaluating, and modeling unionability decisions.

Despite the apparent simplicity of the task, the literature exhibits diverse interpretations of what “unionable” means [14, 18], and even domain experts often disagree when assessing the same table pairs [28]. Different frameworks adopt different criteria, from requiring identical schemas [34] to accepting domain-level correspondence [2, 29] to insisting on preserved inter-column relationships [18]. This diversity in definitions raises a question: *how do humans actually make unionability judgments when faced with table pairs?* Studying how humans make unionability decisions, what drives their judgments, where and why they disagree, and how confident they feel, can inform benchmarks that capture this complexity.

Current research on TUS faces several reproducibility challenges. First, existing benchmarks often provide only binary ground-truth labels without capturing the inherent ambiguity in unionability decisions [12, 18, 29, 31, 41]. In many cases, two domain experts may reasonably disagree on whether certain table pairs should be unioned. Second, most evaluation datasets lack the rich contextual information needed to understand why certain judgments are difficult: Are errors due to subtle schema mismatches? Semantic ambiguity? Or fundamental disagreement about what “unionable” means? Third, the human decision-making process is a black box. We lack systematic data on how people evaluate table pairs and how confident they are in their judgments.

We began addressing these challenges in our prior workshop study [28], where we conducted an experimental survey to analyze how individuals judge unionability. We revealed systematic behavioral patterns, such as confidence-accuracy gaps, overconfidence, and decision-time effects that influence human performance in these complex semantic tasks. We also captured disagreement even on seemingly straightforward cases, revealing that “ground truth” itself is often contested. Example 1 shows one table pair from our survey and disagreement among our users.

#### EXAMPLE 1.1. REASONABLE EXPERT DISAGREEMENT

Consider the table pair in Figure 1. When 15 human annotators (more details about their qualifications and background are provided in Section 3) evaluated the pair, they split 8–7 on unionability with 73% average confidence. The disagreement reveals two fundamentally different interpretations:

**Semantic Equivalence Perspective (8 voted “Yes”):** These experts probably focused on conceptual similarity and overlapping domains. One expert with confidence 97% stated: “columns in both tables have the same meaning.” Another with confidence 92% explained: “both tables are union-able because they have the same type of columns for 2 that is continent and language. And the third row is country and city.”

**SQL UNION Requirements Perspective (7 voted “No”):** These experts seemingly applied strict syntactic criteria from database operations. One expert with confidence 95% provided a detailed technical argument: “To perform a union operation on two tables, the following conditions must be met: Same Number of Columns, Same Column Names, Compatible Data Types. Since the column names and their order are different between the two tables, they do not meet the criteria for a union operation.” Another expert with confidence 94% emphasized the structural mismatch: “Tables A and B cannot be union because they have different structures and incompatible data. table A lists countries and their languages, while table B lists cities and their languages.”

Despite offering thorough and confident explanations, the two groups arrived at conflicting decisions. This pattern suggests the importance of benchmarks that reflect not just end decisions, but also the variability in expert perspectives, confidence, and reasoning.

These preliminary findings motivated the development of TUNE (Table UNIONability with human Evaluation), a benchmark that extends our workshop study into a comprehensive resource for reproducible research. TUNE comprises 464 expert judgments over 26 table pairs, capturing not just binary decisions but also confidence scores, decision times, interaction patterns, and free-text explanations. All annotators have technical backgrounds in data science, ensuring informed judgments while preserving natural variation in interpretation. This design allows researchers to study why humans succeed or fail, how their behaviors correlate with accuracy, and how these signals can be harnessed to enhance automated decision-making. Through this extended benchmark and analysis, we address several key questions:

*Human Understanding.* What cognitive and behavioral factors drive human judgments of table unionability?

*Machine Learning Augmentation.* How effectively can ML models predict or improve human decisions using behavioral and contextual features?

*LLM Synergy.* To what extent can large language models emulate or be enhanced by human reasoning in unionability tasks?

*Hybrid Intelligence.* How can combining human insight with algorithmic and LLM reasoning improve the robustness and explainability of data discovery systems?

To answer these questions, we conduct three complementary investigations using TUNE that bridge human and computational approaches to unionability assessment. First, we evaluate how state-of-the-art TUS methods such as SANTOS [18], Starmie [14], and D3L [2], perform on tasks from the survey, revealing where algorithmic predictions align with or diverge from expert judgments. Second, we develop ML models that leverage behavioral indicators (confidence, timing, click patterns) to predict judgment quality, achieving 84% accuracy (+23% over raw human input). Third, we assess LLMs enhanced with human insights, showing that LLMs can successfully harness human wisdom to match majority voting performance. These experiments collectively demonstrate that combining human behavioral patterns with computational methods, whether through ML classifiers or LLM prompting, yields superior performance compared to either approach in isolation. In summary, this work contributes:

- (1) **TUNE Benchmark:** A publicly available human-centered dataset<sup>1</sup> for studying table unionability, integrating human judgments, behavioral metadata, text explanations, and ground-truth annotations.
- (2) **Behavioral Analysis:** A study of human decision-making in TUNE, revealing overconfidence, diverse reasoning strategies, and meaningful behavioral signals (decision time, click patterns, confidence). We show that these cognitive traces capture ambiguity that binary labels alone cannot represent.
- (3) **ML Experimental Studies:** An extensive ML study showing how individual behavioral indicators (decision time, confidence, interaction patterns) and aggregated crowd signals enhance prediction of unionability correctness. Ablation studies show that decision-time features and aggregated human signals are particularly powerful.
- (4) **LLM Performance and Human–Model Synergy:** A systematic evaluation of open source LLMs on TUNE, demonstrating that human-derived signals can enhance model reliability on difficult unionability decisions.

Our paper is organized as follows. Section 2 reviews background on TUS methods and human-in-the-loop systems. Section 3 presents TUNE, including its construction, human signal analysis, and baseline performance of TUS methods. Section 4 describes our experimental methodology exploring single human-in-the-loop, crowd-in-the-loop, and hybrid configurations. Section 5 discusses key findings, future directions and concludes.

<sup>1</sup>[https://github.com/NinaKlimenkova/TUNE\\_Benchmark/tree/main/data](https://github.com/NinaKlimenkova/TUNE_Benchmark/tree/main/data)

## 2 BACKGROUND AND RELATED WORK

In this section, we formalize the unionability decision problem and situate our work within prior research on table union search and human-in-the-loop data integration.

### 2.1 Task Definition

Let  $\mathcal{T}$  be a universe of tables, where each table  $T \in \mathcal{T}$  is a relation with schema  $Sch(T)$  (attribute names and types) and instance  $I(T)$  (its rows). Given a *query* table  $T_q$  and a *candidate* table  $T_c \in \mathcal{T}$ , the table unionability task is to decide whether  $T_q$  and  $T_c$  can be meaningfully combined by row-wise union under a chosen notion of “unionability” (e.g., compatible entity type per row and semantically aligned columns).

In this paper, we do not address the search problem or the challenge of retrieving candidate tables from large repositories. Our focus is solely on the *decision* aspect of TUS: given a specific pair  $(T_q, T_c)$ , how well can methods judge whether the tables are unionable? Accordingly, our benchmark evaluations in Section 3.4 analyze three widely used TUS methods – *Starmie* [14], *SANTOS* [18], and *D3L* [2], as decision functions rather than search mechanisms.

For our experiments in Section 4, we assume a ground truth  $y \in \{0, 1\}$  indicating whether  $(T_q, T_c)$  is unionable. The *unionability decision problem* is to learn a function  $f : \mathcal{T} \times \mathcal{T} \rightarrow [0, 1]$  such that  $f(T_q, T_c)$  approximates the unionability score that  $y = 1$ . In the *table union search* methods setting, the task becomes: given a query table  $T_q$  and a candidate set  $C \subseteq \mathcal{T}$ , produce a ranking over all  $T_c \in C$  according to their unionability scores  $f(T_q, T_c)$ , so that unionable candidates appear higher than non-unionable ones.

Alongside the ground-truth labels, we observe individual human judgments  $y^{(u)}$  for each annotator  $u$  and collect associated behavioral signals. Our goal is to understand how these human-derived signals relate to unionability outcomes and how they can improve automated estimates of  $f(T_q, T_c)$  within the experimental frameworks introduced in Section 4.

### 2.2 Table Union Search Methods

TUS has evolved significantly over the past decade, driven by the proliferation of data lakes. Despite this progress, the field remains characterized by diverse and often conflicting approaches to defining and solving the unionability problem. Below, we review the state-of-the-art methods, how they define table unionability and discuss their advantages and limitations in the context of our work.

**Early Foundations: Schema-Based Approaches.** Traditional database systems define union operations strictly: two tables  $T_q$  and  $T_c$  are unionable only if their schemas  $Sch(T_q)$  and  $Sch(T_c)$  are identical in attribute names, types, and ordering [34]. This definition, rooted in SQL semantics, ensures syntactic compatibility but proves overly restrictive for real-world data discovery scenarios where tables from heterogeneous sources rarely share perfectly aligned schemas. Historically, one can also apply schema matching to identify corresponding attributes to align the schemas [21, 26, 33, 35, 36, 40].

**Relaxing Constraints: Domain-Based Union Search.** Nargesian et al. [29] proposed a domain-based definition of unionability:  $T_q$  and  $T_c$  are unionable if their corresponding attributes belong to

the same underlying semantic domain (e.g., both represent *countries* even if named differently). Under this perspective, unionability is satisfied when *at least one* pair of attributes can be aligned at the domain level, representing a more flexible, semantics-driven interpretation of the task. The system employs a Locality-Sensitive Hashing (LSH) approach to efficiently identify candidate unionable tables from large repositories, achieving scalability while relaxing strict schema requirements. Building on this LSH foundation, Bogatu et al. [2] introduced D3L, recognizing that no single similarity measure adequately captures unionability. D3L constructs separate LSH indexes for five complementary evidence types: attribute name similarity, attribute value overlap, format patterns, word embeddings, and domain distributions, and aggregates these signals through a weighted distance metric learned from labeled data.

**Advantages & Limitations:** Domain-based methods thus enable discovery across heterogeneously named tables and, in multi-evidence frameworks such as D3L, offer robustness by combining complementary signals. However, they still produce deterministic scores without modeling the uncertainty we observe among domain experts. Two tables may share domains yet differ in granularity (see Example 1), abstraction level, or intended use. These factors strongly affect human decisions  $y^{(u)}$  but are not captured by domain-level similarity alone.

**Semantic Enrichment: Contextual Representations.** Fan et al. [14] advanced the notion of unionability by embedding tables into a semantic vector space. Instead of comparing isolated attributes, their method learns contextualized representations that encode relationships among columns and the higher-order structure of entire tables. In this formulation, the unionability score  $f(T_q, T_c)$  emerges from geometric similarity in the learned embedding space rather than explicit domain matching. Following this representation-learning perspective, Hu et al. [17] propose *AutoTUS*, which learns table-level representations that jointly capture local column signals and global table structure for unionability prediction. More recently, Qiu et al. [32] introduce *LIFTus*, an adaptive multi-aspect column representation model for TUS that is designed to better handle non-linguistic and numeric data by allowing each column to exhibit multiple semantic aspects simultaneously.

**Advantages & Limitations:** These contextual and representation-learning methods can better capture semantic compatibility and often align more closely with human intuitions of when tables are related, while pre-trained or learned representations offer the potential to generalize across heterogeneous domains. Despite capturing richer semantics, these methods still face interpretability challenges. The learned embeddings operate as black boxes, making it difficult to understand why certain tables are deemed similar. They continue to produce binary similarity scores without confidence estimates, and may struggle with domain-specific terminology.

**Relationship-Aware Union Search.** Khatiwada et al. [18] introduced SANTOS, which explicitly requires that unionable tables preserve not just domain correspondence but also inter-column relationships. For example, if a table  $T_q$  exhibits a functional dependency such as *Country*  $\rightarrow$  *Capital*, a unionable table  $T_c$  must respect the same dependency. Rather than only aligning attributes, SANTOS aligns semantic constraints within tables using graph-based matching techniques.

*Advantages & Limitations:* This relationship-aware view helps ensure structural coherence, preventing unions that would violate semantic constraints or introduce inconsistencies. At the same time, this stricter definition can be overly restrictive for some use cases and depends on reliably detected relationships, which is challenging in heterogeneous data lakes. Our evaluation of SANTOS on TUNE confirms these limitations, revealing cases where the relationship-preservation requirement excludes tables that human experts nevertheless consider unionable.

The progression from schema-based matching to relationship-aware methods demonstrates the improvement in capturing unionability. Yet all these approaches share fundamental limitations: they produce deterministic binary predictions, offer limited explainability about their reasoning, and cannot account for the context-dependent nature of unionability. The following section summarizes the earlier work and positions it as the foundational step toward the human-centered perspective we advance in this paper.

## 2.3 Cognitive-aware Data Discovery

The broader data discovery literature acknowledges the indispensable role of human judgment. Cognitive aspects of human decision-making have been examined in related data integration tasks [22, 24, 43], most notably in schema and entity matching. Prior work [23, 37, 38] demonstrate that humans exhibit overconfidence, inconsistent evaluation strategies, and characteristic decision-time patterns, and that some annotators can be systematically identified as more reliable experts. However, these studies focus on matching and alignment tasks, and to the best of our knowledge, *no prior work has applied a similar cognitive, behavior-aware analysis to table unionability itself*. Existing data discovery systems typically assume that humans serve as reliable validators, but they do not examine whether these judgments reflect internal consistency, stable decision criteria, or aligned interpretations of what “unionable” means in practice. This gap is particularly important because unionability is not solely a structural property of tables, it involves interpretation, context, and semantic reasoning that may vary across individuals, as we have already discussed in previous sections.

Complementary to this cognitive line of work, several recent papers critically examine how TUS and related data discovery tasks are benchmarked. Pal et al. [30] propose a generative framework for constructing TUS benchmarks, showing that LLM-generated benchmarks can be substantially more challenging than existing hand-curated ones and reveal different failure modes of state-of-the-art methods. Srinivas et al. [41] introduce LakeBench, a suite of benchmarks for data discovery over data lakes that span unionable, joinable, and subset-related tables, and demonstrate that current tabular foundation models still perform far from satisfactorily on these tasks. Most recently, Boutaleb et al. [3] conduct a systematic re-evaluation of prominent TUS benchmarks and show that simple baselines can achieve surprisingly strong results due to dataset-specific artifacts, arguing for more realistic and carefully designed evaluation resources. While these efforts highlight important limitations of existing benchmarks and call for richer evaluations, they still treat unionability labels as fixed ground truth. They do not analyze the human decision processes that generate those labels.

## 2.4 Preliminary Work

Our earlier workshop paper [28] took a first step toward addressing this gap through an experimental study of unionability judgments. We observed systematic confidence–accuracy mismatches: mean confidence remained high (around 74%–79%), while actual accuracy was noticeably lower (59%–66%). Behavioral traces such as decision time and click patterns emerged as strong predictors of judgment quality—for example, multi-click responses were substantially more accurate than single-click ones—indicating that humans themselves struggle with certain types of unionability decisions and that this struggle is reflected in their interaction patterns.

Building on these observations, we trained traditional ML models on behavioral features (confidence, decision time, clicks, demographics) and achieved an average accuracy improvement of roughly 25 percentage points over raw human labels, with decision-time features alone providing much of the gain. We also examined LLMs in this setting and found that, while they did not consistently outperform humans on their own, they benefited from structured human signals: when provided with human consensus and meta-cognitive metrics in-context, Llama-3.3 70B improved from 59.4% to 75.0%, matching human majority voting.

LLMs add another dimension to the study of unionability [15, 25, 27]. They have demonstrated strong general reasoning capabilities and can perform structured data tasks through few-shot prompting and in-context learning [5]. Specifically, ALT-GEN [31] explored LLMs for table union search using natural-language representations of tables, but did not incorporate human behavior or examine human–LLM synergies. In contrast, our workshop study found that LLMs alone do not consistently outperform humans on unionability decisions, yet they benefit from structured human signals. When provided with human consensus and meta-cognitive metrics (decision time, confidence levels) through in-context learning, LLM (specifically Llama-3.3 70B) accuracy improved from 59.4% to 75.0%, matching the performance of human majority voting. This suggests that combining human judgment patterns with automated methods can yield better outcomes than either approach in isolation, a hypothesis we investigate further in our analysis.

These results highlighted the need for a scalable, behavior-rich benchmark capturing how humans interpret unionability, which is a motivation that directly led to the development of TUNE.

## 3 THE TUNE BENCHMARK

We introduce TUNE (Table UNIONability with human Evaluation), a benchmark combining human judgment and TUS methods predictions over table unionability decisions. It enables analysis of agreement, reasoning patterns, and model-human alignment.

### 3.1 TUNE Overview

TUNE builds upon the TUS benchmark infrastructure, specifically leveraging table pairs from the UGEN benchmarks introduced by Pal et al. [30]. We selected 26 table pairs representing both unionable and non-unionable configurations, extracted from publicly available benchmark ground truth files<sup>2</sup>. These pairs were distributed across four survey versions (V1–V4) and strategically selected to ensure each version had an equal level of difficulty. We deployed

<sup>2</sup><https://github.com/northeastern-datalab/gen/blob/main/data>

Statistic	Value
Table pairs	26
Survey versions	4 (V1–V4)
Unionability questions	32
Expert annotators	58
Total judgments	464
Judgments per pair (avg.)	$\approx 14.5$
Average accuracy (single human)	61.2%
Average accuracy (majority)	75.2%
Average confidence (per version)	71–80%
Average decision time (per question)	85 s
Average click count (per question)	2.2 clicks

**Table 1: Key statistics of the TUNE benchmark.**

a structured online survey through Qualtrics<sup>3</sup>, collecting 464 individual assessments from 58 authenticated participants, all students from our institution with backgrounds in Data Science, Computer Science, Artificial Intelligence, or related fields. Each participant was randomly assigned to one survey version and answered 8 unionability questions, with each question split across two pages: the first capturing the unionability judgment along with behavioral metadata (decision time, click patterns), and the second collecting confidence ratings (0-100 scale via slider) and option to leave explanations about their choice. Key dataset statistics, including counts of tables, judgments, and aggregate human performance, are summarized in Table 1. After completing the main unionability questions, participants also answered a short post-survey module demonstrating their conceptual understanding of unionability (e.g., how they would define unionability and which criteria they considered important). For detailed methodology on survey design, participant recruitment, quality control procedures, and behavioral analysis we refer readers to our repository<sup>4</sup>.

### 3.2 TUNE Composition

The TUNE benchmark consolidates unstructured data into a unified, analysis-ready dataset that captures not only final unionability decisions but also the process by which humans arrive at those decisions. We provide 3 primary data artifacts: (1) *a dataset that includes tables for our survey* including the unionability setup and ground truth labels, (2) *a raw Qualtrics export*, containing all unprocessed responses and metadata, and (3) *cleaned dataset* aligning human decisions with table-level identifiers. In addition, we include a detailed feature-engineering specification outlining how behavioral and aggregated human features were derived, along with the resulting feature-engineered dataset used in our experiments. Full details about the structure of each file are available in our repository.<sup>5</sup> Together, these resources provide 4 major classes of information: (1) unionability decisions and confidence scores, (2) behavioral decision-making signals, (3) qualitative explanations revealing reasoning processes, and (4) post-survey questionnaire

responses capturing participant metacognitive reflections. Below, we describe each component.

**Unionability decisions and confidence scores.** Each survey item presents a pair of tables and asks participants to judge whether they are unionable. This binary input forms the core human judgments for the benchmark. Each question was accompanied by a mandatory confidence rating collected via a slider interface initialized at 50 and ranging from 0 to 100. Participants adjusted the slider to indicate their certainty in their yes/no judgment. The confidence patterns captured in TUNE provide a uniquely informative dimension of the benchmark. Across all survey versions, participants reported consistently high confidence, typically averaging between 74% and 79%, despite achieving only 61% accuracy in their unionability judgments. The distribution of confidence scores is heavily right-skewed, with the upper quartile near 90% and the interquartile range spanning approximately 67%–90%[28]. This systematic gap between expressed certainty and actual correctness reflects a stable overconfidence bias, a well-documented phenomenon in human decision-making[39]. By embedding these metacognitive signals directly into the benchmark, TUNE enables analyses that go beyond simple correctness labels. In Section 3.3, we revisit these confidence-accuracy relationships through calibration and resolution analysis, providing an initial view of how human confidence corresponds to the reliability of unionability judgments.

**Behavioral decision-making signals.** Beyond explicit judgments, TUNE captures rich behavioral metadata providing proxy measures for cognitive load, decision difficulty, and reasoning depth. Each unionability judgment page records multiple time-based and interaction-based signals, enabling detailed analysis of how participants approached the task. These include the raw decision time for each question, internally scaled decision times (normalized per participant), and a series of decomposed temporal components such as initial, mid, and ending decision times reflecting different phases of the evaluation process. TUNE also logs click behaviors, including the number of clicks made prior to submission. These data provide rich behavioral indicators, allowing TUNE to support analyses that reach beyond accuracy alone.

**Qualitative explanations revealing reasoning processes.** We explore the text explanations that accompany participants’ unionability judgments as an additional window into their reasoning, describing the heuristics, assumptions, and domain knowledge humans apply when evaluate table pairs. As illustrated in Example 1, two participants may examine the same pair yet rely on different criteria, underscoring the interpretive nature of unionability.

We first examined the overall sentiment of explanations using a DistilBERT model fine-tuned on SST-2<sup>6</sup>. Our analysis shows that explanations skew strongly negative (285 negative vs. 102 positive), with average correctness slightly higher for negative explanations (64%) than for positive ones (56%). When sentiment is considered together with the ground truth, a clearer pattern emerges: positive explanations are more often correct for “yes” judgments (75.4% accuracy), while negative explanations are more often correct for “no”

<sup>3</sup>[https://wpi.qualtrics.com/jfe/form/SV\\_8jqmFQVl43NcHci](https://wpi.qualtrics.com/jfe/form/SV_8jqmFQVl43NcHci)

<sup>4</sup>[https://github.com/NinaKlimenkova/TUNE\\_Benchmark](https://github.com/NinaKlimenkova/TUNE_Benchmark)

<sup>5</sup>[https://github.com/NinaKlimenkova/TUNE\\_Benchmark/tree/main/data](https://github.com/NinaKlimenkova/TUNE_Benchmark/tree/main/data)

<sup>6</sup><https://huggingface.co/distilbert/distilbert-base-uncased-finetuned-sst-2-english>

judgments (72.1%). In other words, positive framing tends to accompany correctly identified unionable tables, and negative framing tends to accompany correctly identified incompatible tables.

We also consider how specific word choices relate to accuracy. Explanations that used relational terms such as “related”, “similar”, “overlap”, “common” or “different” were more accurate (66.8%) than those that did not (57.2%), it might suggest that attention to structural or content relationships aligns well with the semantics of TUNE. In contrast, explanations invoking more categorical notions such as “concept” or “type” were less accurate (56.9%) than those without these terms (63.0%), implying that very general categorization may be a less reliable reasoning mode for this task. Overall, these patterns provide initial observations on how different forms of reasoning relate to table unionability in our benchmark. By embedding natural-language justifications directly into the benchmark, the TUNE becomes a semantically enriched resource; some more details can be found in our technical report<sup>7</sup>.

**Post-survey questionnaire response.** The post-survey question asked participants to reflect on what makes tables unionable. They were shown a small set of example table pairs and asked to select one or more union-defining criteria (e.g., common schema, overlap, same domain, same concept) or to mark the pair as not unionable. In this limited sample, participants were generally accurate on clearly unionable examples (about 89%) and less accurate on non-unionable ones (about 62%), hinting that recognizing “good” unions may be easier than ruling out borderline or partially overlapping cases. Among correct responses on unionable pairs, options such as “common schema”, “same domain”, and “same concept” were selected most often (100, 97, and 86 selections, respectively), indicating that schema alignment and domain-level consistency are salient cues for many respondents. Given the small number of items and the exploratory nature of this module, these patterns should be viewed as anecdotal, but they provide a qualitative complement to the in-task behavioral signals captured in TUNE.

Having outlined the benchmark components, we now turn from describing TUNE to examining how humans performed on it. The next section analyzes raw human performance, including a top-10 human scenario that simulates how the benchmark behaves when relying only on the strongest individual judgments.

### 3.3 Understanding Humans in TUNE

To understand human capability within TUNE, we analyze how participants performed across versions and how performance changes under selective participation. We further consider a top-10 human scenario, where both majority-vote accuracy and individual average accuracy are computed exclusively from the ten highest-performing participants in each version. This setup does not aim to identify individuals, but rather to simulate how the benchmark would behave if labels were derived only from a curated subset of strong human contributors. It therefore provides a practical upper bound on human annotation quality.

Intuitively, across all versions, selective aggregation of strong annotators improves over average individual performance (Table 2). Raw single-human accuracy ranges from 58–70%, while restricting

Survey Version	Single Human	Top-10 Single	Raw Maj. Voting	Top-10 Maj. Voting
V1	70%	80%	100%	100%
V2	58%	63.70%	50%	75%
V3	58%	68.70%	88%	87.50%
V4	59%	66.20%	63%	75%
Avg	61.25%	69.65%	75.25%	84.37%

Table 2: Raw vs. Top-10 human accuracy comparison.

to the top 10 individuals raises this to 63.7–80%. Majority voting provides the largest gains: raw majority accuracy varies from 50–100%, but top-10 majority voting is consistently higher, between 75% and 100% (with V1 reaching 100%). These patterns indicate that carefully selecting and aggregating high-performing annotators yields a more stable human upper bound for unionability labels.

Understanding the quality of human confidence is essential for modeling human reliability in table unionability tasks. Following metacognitive measurement frameworks from prior work in human-in-the-loop data integration [37], we evaluate two key dimensions of judgment quality: *calibration* and *resolution*.

Calibration measures how well participants’ stated confidence aligns with their actual correctness. A well-calibrated participant reports confidence values that match their empirical accuracy. For example, consistently being correct 70% of the time when expressing 70% confidence; if the same person were only 60% correct at 70% confidence, this would indicate overconfidence, whereas 80% correctness would indicate underconfidence. Formally, following [37], we define calibration for a group  $g$  (e.g., a survey version) as:

$$\text{Cal}(g) = \bar{c}_g - \text{Acc}_g,$$

where  $\bar{c}_g$  is the mean reported confidence and  $\text{Acc}_g$  is the empirical accuracy. Values near zero indicate better calibration; positive values indicate overconfidence and negative values underconfidence.

Resolution evaluates how effectively confidence differentiates between easy and hard items. A participant with higher resolution tends to report higher confidence on items they answer correctly and lower confidence on items they answer incorrectly. Following [37], we measure resolution for a group  $g$  using Goodman–Kruskal’s rank correlation:

$$\text{Res}(g) = \gamma(\mathbf{c}_g, \mathbf{y}_g),$$

where  $\mathbf{c}_g$  is the vector of confidence ratings and  $\mathbf{y}_g$  is the corresponding vector of correctness indicators. Larger positive values indicate that confidence tends to increase with correctness; values near zero indicate little discrimination, and negative values suggest that higher confidence is, on average, associated with errors.

To examine these properties in TUNE, we compute calibration and resolution separately for “yes, these tables are unionable” and “no, these tables are not unionable” responses for each survey version. Table 3 summarizes the results. Calibration scores are positive (0.110–0.238), indicating a general tendency toward overconfidence: reported confidence is, on average, higher than empirical accuracy for both positive and negative judgments. This tendency is weakest

<sup>7</sup>[https://github.com/NinaKlimenkova/TUNE\\_Benchmark/blob/main/Technical\\_Report.pdf](https://github.com/NinaKlimenkova/TUNE_Benchmark/blob/main/Technical_Report.pdf)

Survey Version	Calibration		Resolution	
	Yes	No	Yes	No
V1	<b>0.138</b>	<b>0.110</b>	0.037	0.196
V2	0.205	0.238	0.018	-0.278
V3	0.142	0.118	<b>0.391</b>	-0.026
V4	0.178	0.194	0.079	<b>0.478</b>

**Table 3: Calibration and resolution metrics by response type.**

for V1 “no” responses (0.110) and somewhat stronger for V2 “no” responses (0.238).

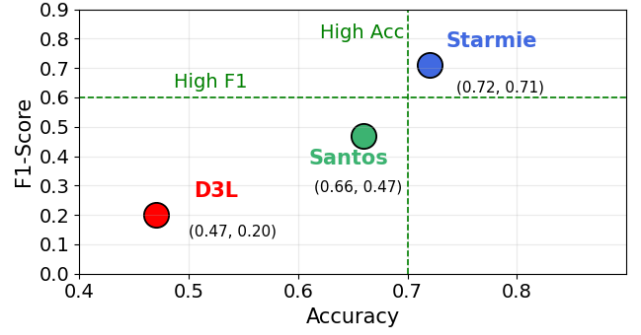
Resolution values are more heterogeneous. For positive responses, they range from 0.018 to 0.391; for negative responses, from -0.278 to 0.478. Most conditions exhibit small to moderate positive resolution, suggesting that confidence carries some information about correctness, with stronger discrimination in V3 “yes” and V4 “no” responses. At the same time, near-zero and negative values (e.g., V2 “no”) indicate that, in some settings, confidence does not reliably separate correct from incorrect judgments, particularly for non-unionable decisions. Overall, these patterns suggest that human judgments in TUNE provide useful but imperfect reliability signals: confidence tends to be somewhat higher than warranted by accuracy, and its discriminative value varies across versions and answer types. Having characterized these metacognitive patterns, we next examine how automated TUS methods behave on the same task.

### 3.4 Benchmarking TUS Methods over TUNE

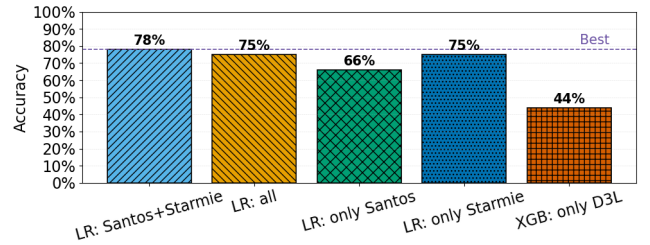
After analyzing human input, we asked ourselves: *how would state-of-the-art TUS methods handle the task that we provided for humans?* To answer this question, we evaluated three representative TUS methods on the same table pairs used in our human survey: Starmie [14], SANTOS [18], and D3L [2]. These methods represent different approaches to TUS (see Section 2) and by benchmarking these methods on TUNE, we can assess both their absolute performance and their potential to complement human judgment.

**TUS performance on TUNE.** To directly compare automated and human decision-making, we evaluated each TUS method on the exact table pairs shown to survey participants, holding conditions constant across all versions. The resulting performance landscape, illustrated in Figure 2, highlights clear differences in method capability. To guide interpretation, we included heuristic “high accuracy” and “high F1” cutoffs in the plot, chosen solely to illustrate the relative positioning of the methods within the performance space.

Consistent with the literature, Starmie lies in the upper-right region (0.72 accuracy, 0.71 F1), indicating relatively strong and balanced performance. SANTOS occupies a middle position (0.66 accuracy, 0.47 F1) showing reasonable capability, while D3L falls in the lower-left (0.47 accuracy, 0.20 F1) suggesting that its multi-evidence approach does not translate effectively to this particular task. In this setting, methods based on contextual semantic representations (Starmie) appear better aligned with TUNE than approaches relying mainly on data lake statistics or multi-evidence similarity. A more detailed analysis of the raw TUS methods performance is provided in our technical report.



**Figure 2: TUS Methods Performance Landscape on TUNE.** Each point represents a method’s average performance across all survey versions, where the x-axis denotes accuracy and the y-axis denotes F1-score.



**Figure 3: Accuracy of ML models using different combinations of TUS method features.** The dashed line marks the best-performing configuration.

The variability observed across TUS methods highlights both their potential and their shortcomings. Motivated by these findings, we enhance TUNE with their similarity scores and proceed to design a series of human-centered experiments that investigate how automated signals, human behavior, and LLM reasoning can be integrated for more robust unionability assessment.

**Combining TUS scores with ML models.** Given the varying performance profiles of different TUS methods, we investigated whether combining their signals could improve overall performance beyond what any single method achieves. We trained ML classifiers such as Logistic Regression (LR) [11], K-Nearest Neighbors (KNN) [10], Random Forest (RF) [4], and XGBoost (XGB) [7], using the scores from Starmie, SANTOS, and D3L as input features. We employed a Leave-One-Version-Out (LOVO) cross-validation strategy (see more details in Section 4) to evaluate model generalization across all survey versions. For each classifier, we conducted an ablation study to evaluate different feature combinations: using all three TUS methods together, each method individually (Starmie only, SANTOS only, D3L only), and selective combinations excluding specific methods (e.g., SANTOS+Starmie). We report results using the best performing model for each configuration, though results were qualitatively similar across classifier types. After conducting a detailed ablation study, we selected the most informative results for visualization. Figure 15 shows the average LOVO accuracy of ML

models using different combinations of TUS scores as features. As single-feature inputs, Starmie reaches 75% accuracy, SANTOS 66%, and D3L only 44%, indicating that Starmie provides the strongest individual signal while D3L is comparatively weak. Using all three methods together yields 75% accuracy, no better than Starmie alone, suggesting that the model largely relies on Starmie and gains little from the other scores. In contrast, the SANTOS+Starmie configuration attains the best performance at 78%, implying that SANTOS contributes complementary signal to Starmie, whereas D3L tends to dilute performance in this setting. More detailed per-version and model-specific results are provided in our repository<sup>8</sup>.

## 4 EXPERIMENTAL METHODOLOGY

Our experimental methodology evaluates how human-generated judgments and TUS-derived features each contribute to producing more accurate and reliable unionability labels over TUNE. Building on the formal decision problem in Section 2, we instantiate four scenarios (S1–S4), where each scenario specifies which inputs are provided to a model when predicting the binary unionability label  $y \in \{0, 1\}$  for a table pair  $(T_q, T_c)$ . We apply both ML classifiers and LLMs under these configurations and compare their behavior across scenarios. In this section, we describe experimental design and each scenario in detail and present the corresponding results.

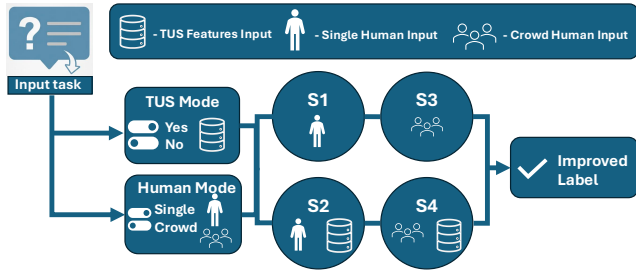


Figure 4: Experimental unionability evaluation pipeline.

### 4.1 Experimental Design Overview

Figure 4 presents our experimental unionability evaluation pipeline. The schema is organized around two binary switches. The *TUS mode* switch determines whether scores from benchmarked TUS methods are available as additional features. The *human mode* switch determines whether the model sees a single annotator (an individual signal with its associated behavioral metadata) or an aggregated representation of the crowd (question-level aggregates such as majority vote and correctness-related statistics). Crossing these switches yields four scenarios: S1 (single human, no TUS), S2 (single human + TUS), S3 (crowd, no TUS), and S4 (crowd + TUS). Each scenario thus corresponds to a specific information configuration for predicting the unionability label of a table pair.

**ML Configuration: Classifiers and Features.** Operationally, each row in our dataset corresponds to a human response to one of the 32 unionability questions (8 questions across 4 versions). For the ML experiments, we start from the feature-engineered TUNE dataset

described in Section 3, where response-level behavioral signals (decision time, click patterns, confidence), user-level metadata, and crowd-level aggregates (e.g., majority vote, correctness counts) have been computed once over the full survey. TUNE is balanced by design, and in every training split the majority-class baseline attains 50% accuracy, providing a simple lower bound. For a given scenario, we select the subset of features consistent with the active modes in Figure 4 (e.g., excluding TUS scores in S1) and train standard classifiers (LR, KNN, RF, and XGB) on this fixed representation. For each scenario we run an ablation over feature groups and classifier types; in the main text we report the best-performing model and feature combination per scenario, while full ablation tables are available in our public repository.<sup>9</sup>

**LLM Configuration: Model Selection and Prompting.** For the LLM experiments, we first evaluated eight open-source models through the Groq API<sup>10</sup> on TUNE questions, collecting for each model a binary unionability decision and a self-reported confidence score on a 0–100 scale. Based on accuracy, calibration, and resolution, we selected Qwen-3-32B [42] as our primary model. We then instantiated the same four information configurations (S1–S4) in prompt form, mirroring the switches in Figure 4. The detailed prompting strategy is available in the technical report. Our goal is not to rank ML classifiers against the LLM, but to observe how each modeling family behaves under the same information scenarios relative to human performance.

**Evaluation protocol and metrics.** For both ML and LLM results, we apply a shared cross-version evaluation protocol (similar to the k-fold cross validation). For ML models, we use a leave-one-version-out (LOVO) scheme over the four survey versions (V1–V4): in each fold, models are trained on three versions and evaluated on the held-out version. For every table pair in the held-out version, we use the benchmark unionability label  $y \in \{0, 1\}$  defined in Section 2 as ground truth, and compare model predictions against this label. For LLMs, each table pair is evaluated with a fresh perspective whether assessing the responses individually or over the aggregated crowd version. The same benchmark label  $y$  is used to compare predictions against our ground truth. This setup mimics deployment to a new batch of unionability questions while preserving the natural grouping induced by the survey design. Final performance metrics, primarily accuracy and F1, and for some analyses calibration and resolution are obtained by averaging over the four folds. Formal definitions of these measures follow the notation in Section 3.3 and are provided in detail in the technical report.

*Note on Scope:* Figures 5–12 report accuracies per survey version. Because each version of TUNE is label-balanced by design (approximately half unionable and half non-unionable questions), micro and weighted F1 scores are numerically very close to accuracy. For clarity, we therefore visualize and discuss *accuracy* for all further scenarios, noting that the corresponding F1 values follow the same pattern and are provided in the technical report. In addition, for each scenario we consider multiple classifiers, feature groups, and ablation variants. Given space constraints, the main text focuses on the most informative configurations. Complete per-model, per-feature, and per-metric results are provided in our technical report.

<sup>8</sup>[https://github.com/NinaKlimenkova/TUNE\\_Benchmark/tree/main/results](https://github.com/NinaKlimenkova/TUNE_Benchmark/tree/main/results)

<sup>9</sup>[https://github.com/NinaKlimenkova/TUNE\\_Benchmark/tree/main/results](https://github.com/NinaKlimenkova/TUNE_Benchmark/tree/main/results)

<sup>10</sup><https://console.groq.com/docs/overview>

We now detail the four scenarios S1-S4 induced by this pipeline, specifying their input configurations in terms of human and TUS signals and examine how these choices shape models' performance and behavior over TUNE.

## 4.2 Scenario 1: Single Human-in-the-loop

**ML setup:** In S1, we model the unionability decision using only the information available from a single annotator at answer time. Each training instance corresponds to an individual response to a TUNE question, and the feature space is restricted to per-user behavioral and demographic signals: the annotator's binary unionability judgment (*SurveyAnswer*), response-level timing and interaction features (e.g., decision time, click counts, click timing), confidence scores, explanation length, and basic user metadata (age, education, English proficiency, major, etc). We explicitly exclude TUS-derived scores, ensuring that the model's predictions in S1 are driven solely by the local signal of a single human plus their behavioral traces. On this feature subset, we train standard classifiers mentioned above and evaluate them under the LOVO protocol described.

**LLM setup:** In S1, the LLM is asked to reassess individual human decisions. Each prompt includes the table-union question, the two tables, and a single human response together with its associated metadata (reported confidence, decision time, and click count). This setup allowed us to observe how the model interprets individual human decisions and how human confidence and behavior affect the model's judgment.

**ML Results:** For the ML setting (Figure 5), we report a representative configuration selected from a broader ablation over classifiers and feature groups: an XGB classifier trained only on the *Decision-Time* feature group. This model reaches an average accuracy of 75.0% across versions, compared to 61.2% for raw human input. On three of the four versions (V1, V3, V4), it substantially improves over raw human input, with 87.5% accuracy versus human accuracies in the 58–70% range. On V2, in contrast, the DecisionTime-only model attains 37.5% while humans achieve 58.0%, indicating that the temporal patterns learned from other versions do not transfer cleanly to this question set. In the full ablation study, some alternative configurations perform better on V2. For example, RF models using only click-based or only confidence features reach about 64% accuracy, but these variants perform worse on other versions and yield lower average accuracy than the DecisionTime-only XGB model. Across models, removing the DECISIONTIME group produces the largest drop in average accuracy (down to 62.5%), whereas dropping other groups (e.g., CLICK, USER-META) has smaller effects and using them alone leads to near-chance performance. These patterns point to decision time as the most informative single-annotator feature in S1. Due to space constraints, we report only representative configurations here; full per-version, per-feature, and per-model results are provided in our technical report.

**LLM Results:** Across all versions, the model reaches an average accuracy of 73.1%, improving on the 61.2% raw human input, with gains observed on each of the four versions. Because in Scenario 1 the LLM is explicitly instructed to output both a binary unionability decision and a 0–100 self-reported confidence score, we can also examine its metacognitive behavior using the calibration and resolution measures introduced in Section 3.3. At the version

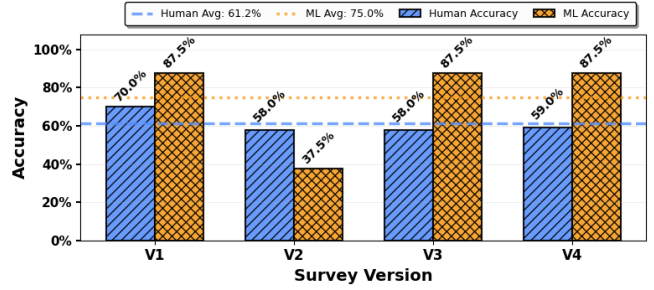


Figure 5: ML Improvement Over Single Human Judgments.

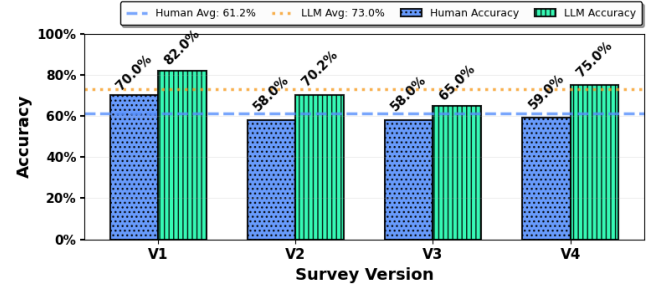


Figure 6: LLM Evaluation of Single Human Judgments.

level, calibration error remains in a moderate range (approximately 0.306–0.423), and resolution values indicate only limited separation between the confidence assigned to correct versus incorrect predictions. In combination with the accuracy results, this suggests that higher LLM confidence tends to be associated with correctness, but the model does not fully exploit the available confidence range to sharply distinguish easy from hard instances.

### Main insights from Scenario 1:

- **Time tells truth:** Decision-time patterns dominate predictive power, while static user demographics contribute less. In our case, how long someone deliberates matters more than who they are.
- **LLMs as refiners:** Given individual votes and meta-data, the LLM seemingly improves noisy human judgments with well-calibrated confidence across versions.
- **Complementary correction:** Both ML and LLM act as effective "second opinions" that boost accuracy over raw votes, suggesting hybrid human-machine workflows outperform either alone.

## 4.3 Scenario 2: Single Human-in-the-loop + TUS

**ML setup:** In S2, we build directly on the single-human setup from S1 but augment the feature space with automated signals from our benchmarked TUS methods. Each training instance is again an individual response to a TUNE question, with the same per-user behavioral and demographic features as in S1. On top of this, we add three features: the unionability scores produced by Starmie, SANTOS, and D3L for the corresponding table pair. The model

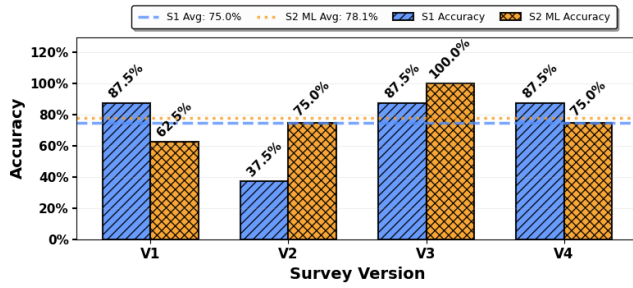


Figure 7: Scenario 1 vs. 2 (ML): Impact of Adding TUS Features to Single Human Input.

therefore learns to refine a single annotator’s decision by jointly exploiting their behavioral traces and the three TUS-derived scores.

**LLM setup:** In addition to the table-union question, the two tables, the human decision, and metadata, each prompt also included similarity scores generated by the strongest TUS signal (the Santos+Starmie combination). The LLM was instructed to integrate both the human inputs and these TUS scores to decide whether to accept or override the human unionability judgment and to report self-confidence in the same manner. This setup allows us to evaluate whether combining an individual human decision with automated TUS signals improves the reliability and accuracy of the model’s judgments. The full prompt templates and wording used in this scenario are provided in the technical report.

**ML Results:** Figures 7 compares S2 to the single-human model performance in S1. Adding TUS scores in S2 yields a modest increase in average accuracy from 75.0% (best S1 model) to 78.1%. The version-wise pattern is mixed: accuracies rise on V2 (37.5% → 75.0%) and V3 (87.5% → 100.0%), but drop on V1 (87.5% → 62.5%) and V4 (87.5% → 75.0%). Overall, performance becomes more even across versions, with all folds now lying between 62.5% and 100.0%. The ablation study for S2 shows that the best configuration is a KNN classifier using only the TUS feature group. Removing TUS features (*drop\_TUS*) consistently reduces average accuracy to around 70% across all 4 classifiers, confirming that the improvements over S1 observed in S2 are largely driven by the additional evidence provided by Starmie, SANTOS, and D3L scores rather than by the single-human features alone.

**LLM Results:** For the LLM (Figure 8), enriching the prompts with TUS scores yields a small average gain over S1: mean accuracy increases from 73.1% (S1) to 74.5% (S2). At the version level, the effect is mixed: accuracy is slightly lower with TUS scores on V1 and V2 (82.0% → 78.9% and 70.2% → 65.4%), but higher on V3 and V4 (65.0% → 71.7% and 75.0% → 82.1%). Based on self-reported confidence, the calibration profile remains in a similar range across both scenarios (approximately 0.30-0.44), while resolution changes more noticeably: it increases in V1 and V4, indicating better separation between correct and incorrect predictions when individual human metadata is combined with TUS scores.

**Main Insights from Scenario 2:**

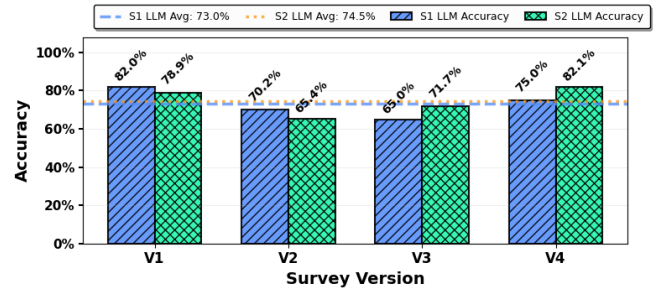


Figure 8: Scenario 1 vs. 2 (LLM): Impact of Adding TUS Scores to Single Human Input

- **TUS dominates:** Once algorithmic scores are available, they carry most predictive weight, dropping them cuts accuracy to 70% while removing behavioral features has minimal impact.
- **ML finds the signal:** The best ML configuration relies substantially on TUS features (KNN + SANTOS/Starmie/D3L), effectively learning to ignore noisier human behavioral traces.
- **LLMs gain selectively:** Enriching prompts with TUS scores yields modest but consistent improvements, with sharper resolution on some versions suggesting better confidence calibration when algorithmic evidence supports human judgment.

#### 4.4 Scenario 3: Crowd-in-the-loop

**ML setup:** In S3, we move fully to a crowd-level representation. Each training instance now corresponds to each TUNE question (one row per question), described by aggregates over all human responses to that question. The feature space consists of 14 crowd features: vote counts and proportions for “yes” and “no”, the binary entropy of the vote distribution, aggregated confidence statistics (overall mean and standard deviation, plus mean confidence conditioned on the majority “yes” and “no” votes), and corresponding mean and standard deviation for decision time and click counts. No TUS-derived scores are included in this scenario. Using these aggregated descriptors, we train the same set of classifiers.

**LLM setup:** Similarly to the ML setup, the LLM evaluated the aggregated human decisions over each question rather than the individual responses. Each prompt contained the table-union question, the tables, and a summary of the human votes, revealing the majority opinion. This setup enabled us to evaluate the LLM’s ability to interpret the crowd consensus and how effectively it utilizes the same to improve its predictions.

**ML Results:** Figure 9 compares S3 to the human majority-vote baseline per question. On average, the majority vote reaches 75.2% accuracy, while the best S3 ML configuration is a LR classifier trained on all crowd-level aggregate features excluding the CONFIDENCE group (*drop-CONFIDENCE*) achieves 78.1%. Version-wise, the model remains close to the majority baseline on V1 and V3, but shows clearer gains on V2 75.0% vs. 50.0% respectively and a small change on V4 62.5% vs. 63.0%. The ablation patterns suggest that

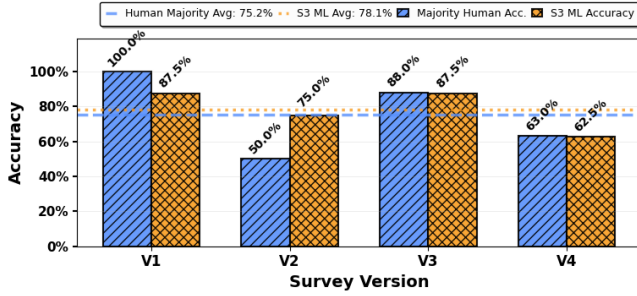


Figure 9: ML Evaluation of Crowd Judgements.

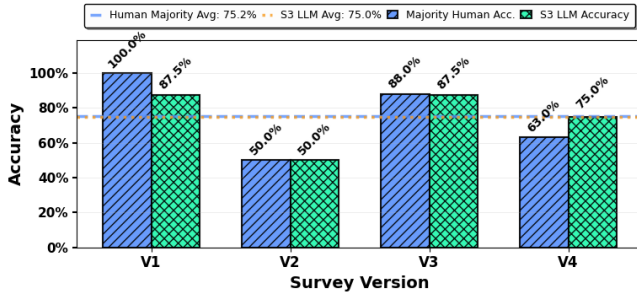


Figure 10: LLM Evaluation of Crowd Judgements.

most of the predictive signal at the crowd level comes from how votes, times, and clicks are distributed across annotators, whereas the additional confidence-aggregate features contribute less and can be omitted without harming performance.

**LLM Results:** With access only to aggregated crowd features, the LLM model attains an average accuracy of 75.0%, essentially matching the human majority (75.2%). Version-wise, its behavior closely tracks the crowd: it reaches 87.5% accuracy on V1 and V3, slightly below the perfect consensus on V1 (100%) and very close to the majority on V3 (88.0%). On V4, the model exceeds the majority baseline (75.0% vs. 63.0%), suggesting that in some cases it can refine noisy or inconsistent crowd judgments. Calibration stays in a moderate range (about 0.27-0.42). Resolution is higher on V1-V3 (around 0.56-0.71) and lower on V4 (0.24), indicating that confidence discriminates best between correct and incorrect predictions when the underlying crowd signal is clearer.

#### Main Insights from Scenario 3:

- **Majority rules, mostly:** Crowd aggregates form a strong 75% baseline. ML refinements exploit disagreement patterns for modest but consistent gains.
- **LLMs mirror the crowd:** When prompted with vote summaries, the LLM essentially tracks majority opinion, occasionally correcting noisy consensus but rarely overriding clear signals.
- **Refinement, not revolution:** Both approaches stabilize rather than transform crowd judgments.

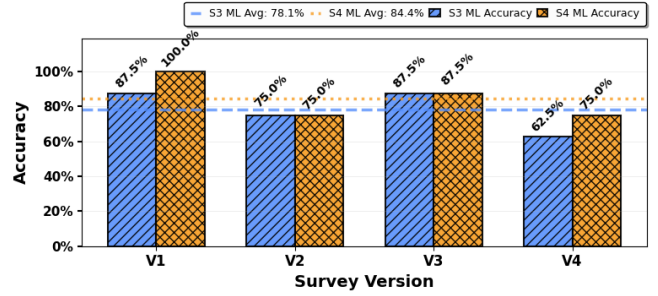


Figure 11: S3 vs. S4 ML Accuracy across survey versions.

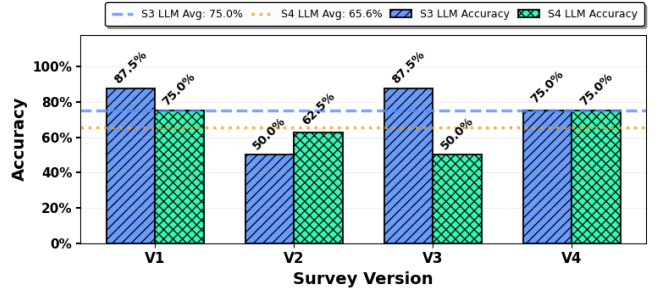


Figure 12: S3 vs. S4 LLM Accuracy across survey versions.

## 4.5 Scenario 4: Crowd-in-the-loop + TUS

**ML setup:** S4 combines the crowd-level representation from S3 with automated evidence from TUS methods. The resulting feature vector jointly captures how the crowd responded to a question and how existing TUS algorithms scored the same pair. Using this combined representation, we train the same set of classifiers under the LOVO protocol, and use ablations to compare models that rely on crowd aggregates, TUS scores, or both.

**LLM setup:** S4 extended the previous aggregated human setup from S3 with the addition of the TUS scores. Each prompt included the table-union question, the tables, a summary of human votes, and the similarity scores from the TUS methods. This setup enabled us to evaluate how the model incorporates crowd consensus and TUS scores, and whether it enhances the model’s judgment.

**ML Results:** Figure 11 shows how performance changes when we add TUS scores to the crowd-level features used in S3. The best ML configuration is a KNN classifier using all feature groups—reaches an average accuracy of 84.4%, the highest across all scenarios, compared to 78.1% in S3. Version-wise, S4 attains 100% accuracy on V1, improves V4 from 62.5% to 75.0%, and matches S3 on V2 and V3 (75.0% and 87.5%). Ablation results indicate that this gain is mainly driven by the interaction between TUS scores and crowd vote statistics: removing the TUS group or vote-count features reduces average accuracy to about 65.6%, while dropping entropy or confidence aggregates has little effect. Overall, in the crowd+TUS setting, most predictive signal comes from the combination of vote distributions and Starmie/SANTOS/D3L scores, with timing and click-based aggregates playing a secondary role.

**LLM Results:** As shown in Figure 12, this setup leads to a drop in average accuracy from 75.0% (S3) to 65.6%. Version-wise, the model

reaches 75.0% on V1 and V4, improves from 50.0% to 62.5% on V2, but falls from 87.5% to 50.0% on V3, indicating that combining crowd summaries with TUS scores is not uniformly beneficial. Calibration remains similar to S3 (around 0.29–0.38), while resolution becomes more uneven: very low on V1 (0.07), higher on V2 and V4 (0.41 and 0.22), and highest on V3 (0.69). This suggests that TUS signals sometimes sharpen confidence separation (especially on V2–V3) but do not consistently translate into accuracy gains.

#### Main Insights from Scenario 4:

- **ML’s peak performance:** Combining crowds and TUS achieves best performance at 84.4%, both vote distributions and TUS scores are essential, removing either drops accuracy to mid-60s.
- **Synergy requires structure:** ML models effectively integrate heterogeneous signals through learned feature weights; entropy and confidence aggregates prove surprisingly dispensable.
- **LLMs struggle with conflict:** Adding TUS scores degrades LLM performance, suggesting prompting-based approaches lack robust mechanisms for reconciling contradictory evidence.

## 5 DISCUSSION AND KEY TAKEAWAYS

In this section, we summarize our main findings and highlight their implications for table unionability and data discovery.

### 5.1 Key Observations

Our results show that human and TUS signals provide complementary yet imperfect evidence about table unionability. Individual human judgments on TUNE are informative yet noisy: average accuracy is moderate, overconfidence is common, and decision-time and confidence patterns vary across versions. Majority voting, especially over stronger annotators, substantially improves reliability, but disagreement and ambiguity persist even on simple-looking cases, reinforcing the view of unionability labels as outcomes of a structured but imperfect decision process. On the automated side, Starmie is the strongest of the three TUS methods, SANTOS adds complementary signal, and D3L is less effective on TUNE. When used as features in ML models, Starmie+SANTOS generally outperform either alone, indicating that combining semantic and relational evidence is more informative than relying on a single score. The four experimental scenarios illustrate how different information configurations can improve labels. With only a single human and behavioral traces (S1), ML models and the LLM act as useful “second opinions”, mainly leveraging decision-time structure. Adding TUS scores on top of this (S2) shifts the signal toward TUS features, with behavioral cues playing a supporting role. At the crowd level (S3), majority voting already forms a strong baseline, with ML and LLM models providing modest refinements. Combining crowd aggregates with TUS scores (S4) yields the strongest ML performance across all settings, while LLM performance becomes more mixed, showing that hybrids that are straightforward for feature-based models are not automatically beneficial in prompt-based setups. Overall, the most reliable unionability assessments arise

when structured human behavior and TUS scores are deliberately combined.

### 5.2 Limitations

Our study has several limitations. First, TUNE currently focuses on a relatively small, curated set of table pairs and a single institutional population of annotators, which may limit the diversity of unionability interpretations we observe. Second, we benchmark a specific set of TUS methods and a specific set of LLM configurations, so our conclusions about strongest signals are contingent on these choices and on our particular feature engineering and prompting designs. Third, we evaluate decision models on pre-selected pairs rather than in a full retrieval setting over large data lakes, which means that our results speak to unionability *assessment* more than to end-to-end table discovery. These constraints should be kept in mind when extrapolating our findings to broader deployments.

### 5.3 Future Directions

The experimental scenarios point to several directions for designing human-centered unionability workflows and data discovery systems. A natural path forward is to develop hybrid architectures in which feature-based models and TUS scores identify high-confidence cases, while ambiguous table pairs are routed to LLMs or human annotators with carefully structured prompts and explanations. Active-learning and adaptive-sampling strategies could further exploit behavioral signals to focus additional annotation effort on borderline cases. More broadly, our findings motivate evaluation protocols and metrics that explicitly account for disagreement and ambiguity, for example, by using distributional or multi-perspective labels and uncertainty-aware measures, rather than treating variability in human judgments as pure noise.

## 6 CONCLUSION

Table unionability is often treated as if it had clear ground truth. Our study shows it is inherently interpretive: judgments reflect diverse reasoning strategies, varying expertise, and contextual views of what makes tables unionable. Through TUNE, we provide a benchmark that captures not only final unionability decisions but also the cognitive and behavioral context in which those decisions are made. Across 4 experimental scenarios, we show how human-derived behavioral information and TUS methods can be combined: ML models trained on behavioral features and TUS scores reach up to 84% accuracy in the crowd+TUS setting, improving over both raw human judgments and standalone TUS methods, while LLMs can be sensitive to conflicting human and algorithmic signals.

These findings suggest that data discovery systems should favor hybrid architectures over purely human- or model-centric designs. Behavioral signals such as decision time, confidence, and interaction patterns provide valuable information for quality control and uncertainty estimation, while TUS methods contribute structured evidence about schema and content similarity. By making behavioral metadata, confidence scores, textual explanations, and post-survey reflections available alongside traditional labels, TUNE enables researchers to study not just *what* humans decide, but *how* and *why*, supporting more transparent and human-aligned unionability assessments.

## REFERENCES

- [1] Ziawasch Abedjan, Mahdi Esmailoghli, and Sainyam Galhotra. 2025. Data Discovery in Data Lakes: Operations, Indexes, Systems. *Proc. VLDB Endow.* 18, 12 (Aug. 2025), 5455–5459. <https://doi.org/10.14778/3750601.3750694>
- [2] Alex Bogatu, Alvaro A. A. Fernandes, Norman W. Paton, and Nikolaos Konstantinou. 2020. Dataset Discovery in Data Lakes. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. 709–720. <https://doi.org/10.1109/ICDE48307.2020.00067>
- [3] Allaa Boutaleb, Bernd Amann, Hubert Naacke, and Rafael Angarita. 2025. Something’s Fishy In The Data Lake: A Critical Re-evaluation of Table Union Search Benchmarks. <https://doi.org/10.48550/arXiv.2505.21329>
- [4] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32. <https://doi.org/10.1023/A:1010933404324>
- [5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (Vancouver, BC, Canada) (NIPS ’20)*. Curran Associates Inc., Red Hook, NY, USA, Article 159, 25 pages.
- [6] Adriane Chapman, Elena Simperl, Laura Koesten, Natalia Konstantinova, Sergej Sizov, Jon Hare, and Elena Simperl. 2020. Dataset Search: A Survey. *The VLDB Journal* 29, 1 (2020), 251–272. <https://doi.org/10.1007/s00778-019-00564-x>
- [7] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 785–794. <https://doi.org/10.1145/2939672.2939785>
- [8] Martin Pekár Christensen, Aristotelis Leventidis, Matteo Lissandrini, Laura Di Rocco, Renée J. Miller, and Katja Hose. 2025. Fantastic Tables and Where to Find Them: Table Search in Semantic Data Lakes. In *International Conference on Extending Database Technology*. <https://api.semanticscholar.org/CorpusID:274142197>
- [9] Tianji Cong, Fatemeh Nargesian, and HV Jagadish. 2023. Pylon: Semantic table union search in data lakes. *arXiv preprint arXiv:2301.04901* (2023).
- [10] Thomas M. Cover and Peter E. Hart. 1967. Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory* 13, 1 (1967), 21–27. <https://doi.org/10.1109/TIT.1967.1053964>
- [11] David R. Cox. 1958. The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society: Series B* 20, 2 (1958), 215–242.
- [12] Yuhao Deng, Chengliang Chai, Lei Cao, Qin Yuan, Siyuan Chen, Yanrui Yu, Zhaoze Sun, Junyi Wang, Jiajun Li, Ziqi Cao, Kaisen Jin, Chi Zhang, Yuqing Jiang, Yuanfang Zhang, Yuping Wang, Ye Yuan, Guoren Wang, and Nan Tang. 2024. LakeBench: A Benchmark for Discovering Joinable and Unionable Tables in Data Lakes. *Proc. VLDB Endow.* 17, 8 (April 2024), 1925–1938. <https://doi.org/10.14778/3659437.3659448>
- [13] Grace Fan, Jin Wang, Yuliang Li, and Renée J. Miller. 2023. Table Discovery in Data Lakes: State-of-the-art and Future Directions. In *Companion of the 2023 International Conference on Management of Data (Seattle, WA, USA) (SIGMOD ’23)*. Association for Computing Machinery, New York, NY, USA, 69–75. <https://doi.org/10.1145/3555041.3589409>
- [14] Grace Fan, Jin Wang, Yuliang Li, Dan Zhang, and Renée J. Miller. 2023. Semantics-Aware Dataset Discovery from Data Lakes with Contextualized Column-Based Representation Learning. *Proc. VLDB Endow.* 16, 7 (March 2023), 1726–1739. <https://doi.org/10.14778/3587136.3587146>
- [15] Juliana Freire, Grace Fan, Benjamin Feuer, Christos Koutras, Yurong Liu, Eduardo Peña, Aécio SR Santos, Cláudio T Silva, and Eden Wu. 2025. Large Language Models for Data Discovery and Integration: Challenges and Opportunities. *IEEE Data Eng. Bull.* 49, 1 (2025), 3–31.
- [16] Xuming Hu, Shen Wang, Xiao Qin, Chuan Lei, Zhengyuan Shen, Christos Faloutsos, Asterios Katsifodimos, George Karypis, Lijie Wen, and Philip S. Yu. 2023. Automatic Table Union Search with Tabular Representation Learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 3786–3800. <https://doi.org/10.18653/v1/2023.findings-acl.233>
- [17] Xuming Hu, Shen Wang, Xiao Qin, Chuan Lei, Zhengyuan Shen, Christos Faloutsos, Asterios Katsifodimos, George Karypis, Lijie Wen, and Philip S. Yu. 2023. Automatic Table Union Search with Tabular Representation Learning. In *Annual Meeting of the Association for Computational Linguistics*. <https://api.semanticscholar.org/CorpusID:259858834>
- [18] Aamod Khatiwada, Grace Fan, Roe Shraga, Zixuan Chen, Wolfgang Gatterbauer, Renée J. Miller, and Mirek Riedewald. 2023. SANTOS: Relationship-based Semantic Table Union Search. *Proc. ACM Manag. Data* 1, 1, Article 9 (May 2023), 25 pages. <https://doi.org/10.1145/3588689>
- [19] Aamod Khatiwada, Harsha Kokel, Ibrahim Abdelaziz, Subhajit Chaudhury, Julian Dolby, Oktie Hassanzadeh, Zhenhan Huang, Tejaswini Pedapati, Horst Samulowitz, and Kavitha Srinivas. 2025. TabSketch: Sketch-based tabular representation learning for data discovery over data lakes. In *2025 IEEE 41st International Conference on Data Engineering (ICDE)*. IEEE, 1523–1536.
- [20] Aamod Khatiwada, Roe Shraga, and Renée J. Miller. 2025. Diverse Unionable Tuple Search: Novelty-Driven Discovery in Data Lakes [Technical Report]. *arXiv preprint arXiv:2509.01012* (2025).
- [21] Christos Koutras, George Siachamis, Andra Ionescu, Kyriakos Psarakis, Jerry Brons, Marios Fragkoulis, Christoph Lof, Angela Bonifati, and Asterios Katsifodimos. 2021. Valentine: Evaluating matching techniques for dataset discovery. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE, 468–479.
- [22] Guoliang Li. 2017. Human-in-the-loop data integration. *Proceedings of the VLDB Endowment* 10, 12 (2017), 2006–2017.
- [23] Guoliang Li, Jiannan Wang, Yudian Zheng, and Michael J. Franklin. 2016. Crowdsourced Data Management: A Survey. *IEEE Transactions on Knowledge and Data Engineering* 28, 9 (2016), 2296–2319. <https://doi.org/10.1109/TKDE.2016.2535242>
- [24] Huan Yu Li, Zlatan Dragisic, Daniel Faria, Valentina Ivanova, Ernesto Jiménez-Ruiz, Patrick Lambrix, and Catia Pesquita. 2019. User validation in ontology alignment: functional assessment and impact. *The Knowledge Engineering Review* 34 (2019).
- [25] Kaiyu Li, Zhongxin Hu, Yuxin Gao, and Yuyang Wu. [n.d.]. DeepSearch: LLM-powered Data Acquisition for Machine Learning. *Proceedings of the VLDB Endowment*. ISSN 2150 ([n.d.]), 8097.
- [26] Yurong Liu, Eduardo Pena, Aécio Santos, Eden Wu, and Juliana Freire. 2024. Magnet: Combining small and large language models for schema matching. *arXiv preprint arXiv:2412.08194* (2024).
- [27] Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeet Singh Meena, Aryan Prakhhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. 2024. Discoverybench: Towards data-driven discovery with large language models. *arXiv preprint arXiv:2407.01725* (2024).
- [28] Sreeram Marimuthu, Nina Klimenkova, and Roe Shraga. 2025. Humans, Machine Learning, and Language Models in Union: A Cognitive Study on Table Unionability. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics (Intercontinental Berlin, Berlin, Germany) (HILDA ’25)*. Association for Computing Machinery, New York, NY, USA, Article 6, 7 pages. <https://doi.org/10.1145/3736733.3736740>
- [29] Fatemeh Nargesian, Erkang Zhu, Ken Q. Pu, and Renée J. Miller. 2018. Table union search on open data. *Proc. VLDB Endow.* 11, 7 (March 2018), 813–825. <https://doi.org/10.14778/3192965.3192973>
- [30] Koyena Pal, Aamod Khatiwada, Roe Shraga, and Renée J. Miller. 2023. Generative Benchmark Creation for Table Union Search. <https://doi.org/10.48550/arXiv.2308.03883>
- [31] Koyena Pal, Aamod Khatiwada, Roe Shraga, and Renée J. Miller. 2024. ALT-GEN: Benchmarking Table Union Search using Large Language Models. In *VLDB Workshops*. <https://vldb.org/workshops/2024/proceedings/TaDA/TaDA.3.pdf>
- [32] Ermu Qiu, Jun Gao, Yaofeng Tu, and Jingru Yang. 2025. LIFTus: An Adaptive Multi-Aspect Column Representation Learning for Table Union Search. *2025 IEEE 41st International Conference on Data Engineering (ICDE) (2025)*, 2174–2187. <https://api.semanticscholar.org/CorpusID:280693810>
- [33] Erhard Rahm and Philip Bernstein. 2001. A Survey of Approaches to Automatic Schema Matching. *VLDB J.* 10 (12 2001), 334–350. <https://doi.org/10.1007/s007780100057>
- [34] Raghu Ramakrishnan and Johannes Gehrke. 2003. *Database Management Systems* (3 ed.). McGraw-Hill.
- [35] Nabeel Seedat and Mihaela van der Schaar. [n.d.]. Matchmaker: Self-Improving Compositional LLM Programs for Table Schema Matching. In *NeurIPS 2024 Third Table Representation Learning Workshop*.
- [36] Eitam Sheerit, Menachem Brief, Moshik Mishaali, and Oren Elisha. 2024. Rematch: Retrieval enhanced schema matching with llms. *arXiv preprint arXiv:2403.01567* (2024).
- [37] Roe Shraga, Ofra Amir, and Avigdor Gal. 2021. Learning to characterize matching experts. In *Proceedings - 2021 IEEE 37th International Conference on Data Engineering, ICDE 2021 (Proceedings - International Conference on Data Engineering)*. 1236–1247. <https://doi.org/10.1109/ICDE51399.2021.00111> Publisher Copyright: © 2021 IEEE.; 37th IEEE International Conference on Data Engineering, ICDE 2021 ; Conference date: 19-04-2021 Through 22-04-2021.
- [38] Roe Shraga and Avigdor Gal. 2022. PoWareMatch: A Quality-aware Deep Learning Approach to Improve Human Schema Matching. *J. Data and Information Quality* 14, 3, Article 16 (May 2022), 27 pages. <https://doi.org/10.1145/3483423>
- [39] Roe Shraga, Avigdor Gal, and Haggai Roitman. 2018. What Type of a Matcher Are You? Coordination of Human and Algorithmic Matchers. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics (Houston, TX, USA) (HILDA ’18)*. Association for Computing Machinery, New York, NY, USA, Article 12, 7 pages. <https://doi.org/10.1145/3209900.3209905>
- [40] Roe Shraga, Avigdor Gal, and Haggai Roitman. 2020. ADnEV: Cross-Domain Schema Matching using Deep Similarity Matrix Adjustment and Evaluation.

*Proceedings of the VLDB Endowment* 13, 9 (2020), 1401–1415. <https://doi.org/10.14778/3397230.3397237>

- [41] Kavitha Srinivas, Julian Dolby, Ibrahim Abdelaziz, Oktie Hassanzadeh, Harsha Kokel, Aamod Khatiwada, Tejaswini Pedapati, Subhajit Chaudhury, and Horst Samulowitz. 2023. LakeBench: Benchmarks for Data Discovery over Data Lakes. arXiv:2307.04217 [cs.DB] <https://arxiv.org/abs/2307.04217>
- [42] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and Zihan Qiu. 2025. Qwen3 Technical Report. <https://doi.org/10.48550/arXiv.2505.09388>
- [43] Chen Zhang, Lei Chen, HV Jagadish, Mengchen Zhang, and Yongxin Tong. 2018. Reducing Uncertainty of Schema Matching via Crowdsourcing with Accuracy Rates. *IEEE Transactions on Knowledge and Data Engineering* (2018).

## A APPENDIX

### A.1 Benchmarking LLM Performance over TUNE

We evaluated 8 open-source large language models available through the Groq API console<sup>11</sup> namely, llama-3.1-8b-instant, llama-3.3-70b-versatile, meta-llama/llama-4-maverick-17b-128e-instruct, meta-llama/llama-4-scout-17b-16e-instruct, gemma2-9b-it, moonshotai/kimi-k2-instruct, qwen/qwen-3-32b, and deepseek-r1-distill-llama-70b. Our primary goal was to identify the best-performing model for our Human-centered experiments. Each model was presented with 32 table unionability questions across our four versions of the main survey (V1-V4), in the same format shown to our human participants. For every question, models were prompted to give (i) a binary judgment on whether the table pair was unionable or not, and (ii) a self-reported confidence score for their decision on a scale of 0-100.

From our results, we calculated the mean accuracy, and mean confidence for all the evaluated models as shown in Figure 13. Three models particularly qwen-3-32b, llama-4-maverick-17b-128e-instruct, and llama-4-scout-17b-16e-instruct all achieved the same accuracy of 0.719, notably qwen-3-32b with the highest mean confidence score of 80.

Given that accuracy alone could not distinguish between these models, we examined their calibration and resolution, to assess both correctness and reliability. Please find the results of the same below:

- **llama-4-maverick-17b:** calibration = 0.509, resolution = 0.129
- **llama-4-scout-17b:** calibration = 0.575, resolution = 0.189
- **qwen-3-32b:** calibration = 0.319, resolution = 0.246

The qwen-3-32b model displayed the lowest calibration error and the highest resolution, indicating more reliable confidence estimates and better separation between correct and incorrect predictions. Based on this combined overall results, qwen-3-32b was selected as our primary model for the human-centered experiments over TUNE.

Based on our benchmarking of the open-source large language models in terms of accuracy, calibration, and resolution, we selected Qwen-3-32b as the primary model for our human-centered experiments. This model was tested across 4 experimental scenarios, acting as a judge over the individual human votes with metadata, aggregated question-wise human votes with metadata, and both these variants combined with our benchmarked TUS methods’ (Starmie + SANTOS) scores. We thoroughly evaluated the model’s performance survey-wise, calculating accuracy and F1 scores. These evaluations allowed us to compare the model’s performance with human judgments and assess its decision-making behavior across different experimental conditions.

### A.2 Benchmarking TUS Methods over TUNE

After analyzing human input, we asked: *how would state-of-the-art TUS methods handle the same task that we provided to humans?* To answer this question, we evaluated three representative TUS methods on the same table pairs used in our human survey: Starmie [14], SANTOS [18], and D3L [2]. These systems instantiate different

<sup>11</sup><https://console.groq.com/docs/overview>

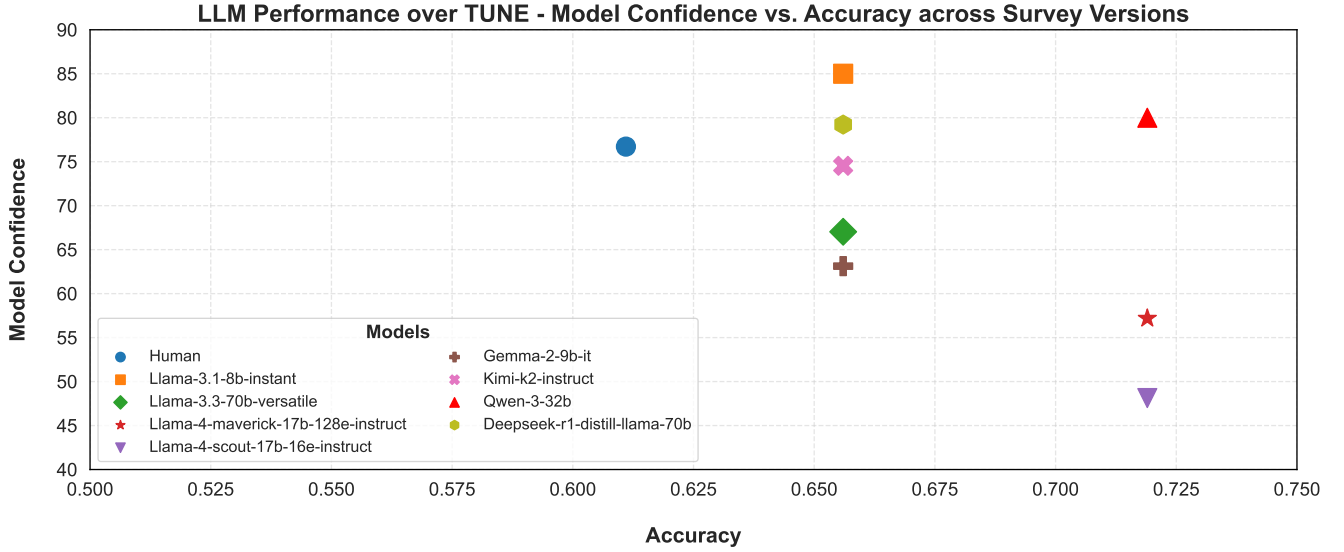


Figure 13: Comparison of Accuracy: Humans vs Different Large Language Models.

notions of unionability (see Section 2), and by benchmarking them on TUNE we can assess both their absolute performance and their potential to complement human judgments.

**TUS performance on TUNE.** To directly compare automated and human decision-making, we evaluated each TUS method on the exact table pairs shown to survey participants, holding conditions constant across all versions. The resulting performance landscape, illustrated in Figure 14, highlights clear differences in method capability. To guide interpretation, we included heuristic “high accuracy” and “high F1” cutoffs in the plot, chosen solely to illustrate the relative positioning of the methods within the performance space.

Consistent with prior reports, Starmie lies in the upper-right region (0.72 accuracy, 0.71 F1), indicating relatively strong and balanced performance. Its scores approach, but do not match, the human majority baseline (75.2%), suggesting that semantic embedding methods can capture much of the signal needed for unionability decisions on TUNE, while still missing some of the nuanced cases that challenge humans. SANTOS occupies a middle position (0.66 accuracy, 0.47 F1): it is clearly better than chance and often identifies unionable pairs correctly, but the gap between accuracy and F1 indicates a less balanced precision–recall profile, likely reflecting a stronger bias toward one class. D3L falls in the lower-left region (0.47 accuracy, 0.20 F1), indicating that its multi-evidence approach (names, values, formats, embeddings, domains) does not transfer well to this particular task and question set. In the TUNE setting, methods based on contextual semantic representations (Starmie) appear better aligned with our benchmark than approaches relying mainly on data lake statistics or heterogeneous heuristic signals. A more detailed analysis of per-version performance and confusion patterns is provided in our technical report.

The variability observed across TUS methods highlights both their potential and their shortcomings. Motivated by these findings,

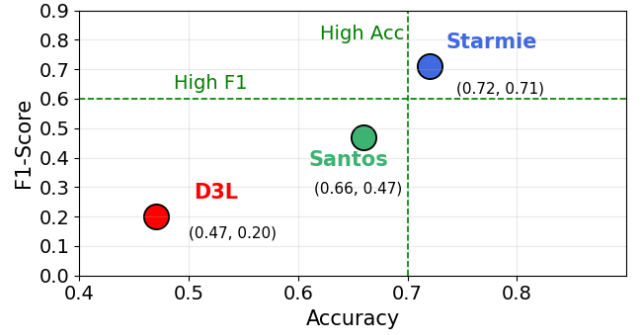
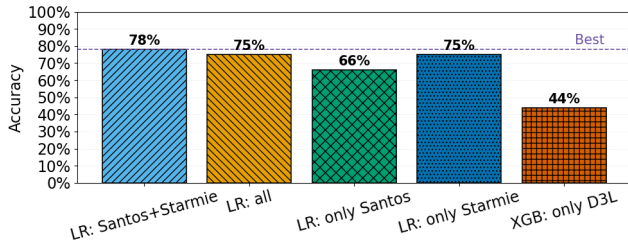


Figure 14: TUS methods performance landscape on TUNE. Each point represents a method’s average performance across all survey versions, where the x-axis denotes accuracy and the y-axis denotes F1-score.

we enhance TUNE with their similarity scores and use them as features in downstream models, enabling a more fine-grained analysis of how automated signals interact with human behavior.

**Combining TUS scores with ML models.** Given the differing performance profiles of Starmie, SANTOS, and D3L, we next examined whether combining their scores through supervised learning could improve unionability prediction beyond what any single method achieves. We trained four standard ML classifiers—logistic regression (LR) [11],  $k$ -nearest neighbors (KNN) [10], random forest (RF) [4], and XGBoost (XGB) [7]—using the outputs of Starmie, SANTOS, and D3L as input features. We employed a leave-one-version-out (LOVO) cross-validation strategy (see Section 4) to evaluate generalization across the four survey versions.



**Figure 15: Accuracy of ML classifiers using different combinations of TUS-method features. The dashed line marks the best-performing configuration.**

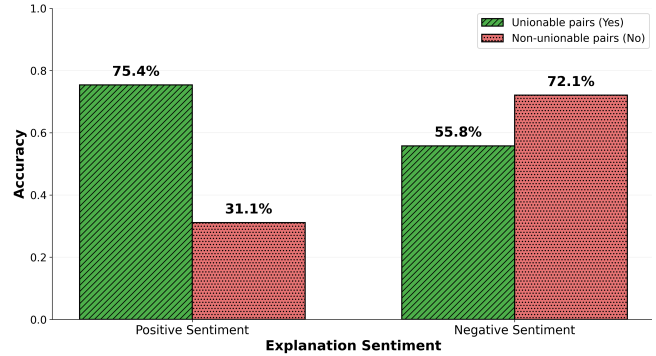
For each classifier, we conducted an ablation study over feature combinations: using all three TUS methods together, each method individually (Starmie-only, SANTOS-only, D3L-only), and selective combinations excluding specific methods (e.g., SANTOS+Starmie). We report in Figure 15 the best-performing configuration per feature set, noting that trends were qualitatively similar across classifier types. As single-feature inputs, Starmie reaches 75% accuracy, SANTOS 66%, and D3L only 44%, confirming that Starmie provides the strongest individual signal while D3L is comparatively weak. Using all three methods together yields 75% accuracy, no better than Starmie alone, suggesting that the model largely relies on Starmie and gains little from the other scores. In contrast, the SANTOS+Starmie configuration attains the best performance at 78%, indicating that SANTOS contributes complementary information to Starmie, whereas D3L tends to dilute performance in this setting.

Overall, these results suggest that while individual TUS scores already carry substantial signal about unionability on TUNE, carefully constructed combinations—particularly Starmie with SANTOS—can modestly improve on the best single method. More detailed

### A.3 Sentiment analysis

To better understand these reasoning patterns, we evaluated the sentiment of each explanation using a DistilBERT model fine-tuned on SST-2<sup>12</sup>, which is well-suited for short English text. We observed that the explanations skew strongly negative: 285 negative vs. 102 positive, with average correctness slightly higher for negative explanations (64%) than positive ones (56%). A clearer picture emerges when sentiment categories are grouped with their corresponding benchmark answers, as illustrated in Figure 16. Positive explanations align strongly with correct “yes” judgments (75.4% accuracy), while negative explanations align strongly with correct “no” judgments (72.1%). This alignment pattern indicates that participants’ framing meaningfully relates to the underlying correctness of their unionability judgments. According to the observation, positive sentiment serves as a heuristic signal for perceived compatibility, while negative sentiment signals perceived incompatibility. The high accuracy rates for aligned combinations (75.4% and 72.1%) suggest that this affective heuristic generally reflects accurate interpretation of table relationships.

<sup>12</sup><https://huggingface.co/distilbert/distilbert-base-uncased-finetuned-sst-2-english>



**Figure 16: Accuracy by Sentiment and Benchmark**

## A.4 Prompt Templates

### A.4.1 Instruction Header (common to all 4 scenarios).

You **MUST** respond **ONLY** in the format: `Accept,<confidence>` or `Reject,<confidence>`. Your answer **MUST** be on a single line. No explanation. No reasoning. No `<think>`. No quotes. No extra words.

### A.4.2 Scenario 1: Single Human-in-the-Loop.

Let’s conduct an experiment. Think carefully and answer only in the specified format. The full question and tables are provided below, followed by the human’s answer and metadata. You are evaluating whether to accept a human’s judgment about a table-unionability using the details below: answer only whether you “accept” or “reject”, along with your confidence level on a scale of 0-100.

**QUESTION:** <Insert question here>

**TABLES:** <Insert tables here>

**HUMAN VOTE & METADATA:**

- vote: “<vote>”
- reported\_confidence(0-100): <conf>
- decision\_time(in sec): <time>
- number\_of\_clicks: <clicks>

### A.4.3 Scenario 2: Single Human-in-the-Loop + TUS Features.

Let’s conduct an experiment. Think carefully and answer only in the specified format. The full question and tables are provided below, followed by the human’s answer, its metadata, and the automated SOTA Table Union Search Algorithms’ similarity scores: Starmie and Santos (range 0.0-1.0). Higher values are stronger evidence of unionability. You are evaluating whether to accept a human’s judgment about a table-unionability using the available information below: answer only whether you “accept” or “reject”, along with your confidence level on a scale of 0-100.

**QUESTION:** <Insert question here>

**TABLES:** <Insert tables here>

**HUMAN VOTE & METADATA:**

- vote: “<vote>”
- reported\_confidence(0-100): <conf>
- decision\_time(in sec): <time>
- number\_of\_clicks: <clicks>

**AUTOMATED TUS SCORES (0.0-1.0):**

- Starmie: <starmie\_score>

- Santos: <santos\_score>

#### A.4.4 Scenario 3: Crowd-in-the-Loop.

Let’s conduct an experiment. Think carefully and answer only in the specified format. You are evaluating whether the tables are unionable based on the aggregated summary of human votes. The full question and tables are provided below, followed by the summary from the human responses. Decide whether the correct judgment is “Yes” (the tables are unionable) or “No” (the tables are not unionable), and provide your confidence level on a scale of 0–100, based on the information below.

**QUESTION:** <Insert question here>

**TABLES:** <Insert tables here>

**GROUP SUMMARY:**

- Count\_Yes: <Count\_Yes>

- Count\_No: <Count\_No>

#### A.4.5 Scenario 4: Crowd-in-the-Loop + TUS Features.

Let’s conduct an experiment. Think carefully and answer only in the specified format. You are evaluating whether the tables are unionable based on the aggregated summary of human votes, with the automated SOTA Table Union Search Algorithms’ similarity scores: Starmie and Santos (range 0.0–1.0). Higher values are stronger evidence of unionability. Decide whether the correct judgment is “Yes” (the tables are unionable) or “No” (the tables are not unionable), and provide your confidence level on a scale of 0–100, based on all the information available below.

**QUESTION:** <Insert question here>

**TABLES:** <Insert tables here>

**GROUP SUMMARY:**

- Count\_Yes: <Count\_Yes>

- Count\_No: <Count\_No>

**AUTOMATED TUS SCORES (0.0–1.0):**

- Starmie: <starmie\_score>

- Santos: <santos\_score>

### A.5 Extended results

SV	ML		LLM	
	Accuracy	F1	Accuracy	F1
1	0.875	0.873	0.820 (+16.6%)	0.818
2	0.375	0.365	0.702 (+21.7%)	0.697
3	0.875	0.873	0.650 (+13.0%)	0.649
4	0.875	0.873	0.750 (+27.3%)	0.749
<b>Avg.</b>	0.75	0.746	0.731 (+19.5%)	0.728

### A.6 Measures definition

Final performance metrics, primarily accuracy and F1, and for some analyses calibration and resolution, are obtained by averaging over the four LOVO folds.

SV	ML		LLM	
	Accuracy	F1	Accuracy	F1
1	0.625	0.563	0.789 (+12.2%)	0.783
2	0.75	0.75	0.654 (+13.4%)	0.653
3	1	1	0.717 (+24.7%)	0.715
4	0.75	0.733	0.821 (+39.3%)	0.821
<b>Avg.</b>	0.781	0.774	0.745 (+22.0%)	0.743

**Table 4: Comparison of Accuracy: ML vs LLM in Single Human-in-the-Loop with TUS Features across survey versions.**

SV	ML		LLM	
	Accuracy	F1	Accuracy	F1
1	0.875	0.873	0.875 (+24.4%)	0.873
2	0.75	0.733	0.500 (-13.3%)	0.500
3	0.875	0.873	0.875 (+52.2%)	0.873
4	0.625	0.619	0.750 (+27.3%)	0.750
<b>Avg.</b>	0.781	0.774	0.750 (+22.7%)	0.749

**Table 5: Comparison of Accuracy: ML vs LLM in Crowd-in-the-Loop across survey versions.**

SV	ML		LLM	
	Accuracy	F1	Accuracy	F1
1	1	1	0.750 (+6.7%)	0.733
2	0.75	0.75	0.625 (+8.3%)	0.619
3	0.875	0.873	0.500 (-13.0%)	0.333
4	0.75	0.75	0.750 (+27.3%)	0.733
<b>Avg.</b>	0.843	0.843	0.656 (+7.4%)	0.605

**Table 6: Comparison of Accuracy: ML vs LLM in Crowd-in-the-Loop with TUS Features across survey versions.**

Formally, let  $\{(y_i, \hat{y}_i, c_i)\}_{i=1}^N$  be the set of  $N$  evaluation instances, where  $y_i \in \{0, 1\}$  is the benchmark unionability label,  $\hat{y}_i \in \{0, 1\}$  is the model prediction, and  $c_i \in [0, 1]$  is the model (or human) confidence mapped to  $[0, 1]$ .

*Accuracy and F1.* Define

$$TP = \sum_{i=1}^N \mathbb{1}[\hat{y}_i = 1 \wedge y_i = 1], \quad FP = \sum_{i=1}^N \mathbb{1}[\hat{y}_i = 1 \wedge y_i = 0],$$

$$\text{FN} = \sum_{i=1}^N \mathbb{1}[\hat{y}_i = 0 \wedge y_i = 1], \quad \text{TN} = \sum_{i=1}^N \mathbb{1}[\hat{y}_i = 0 \wedge y_i = 0].$$

Accuracy is

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{N}.$$

Precision and recall are

$$\text{Prec} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Rec} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

and the F1-score is the harmonic mean

$$\text{F1} = \frac{2 \cdot \text{Prec} \cdot \text{Rec}}{\text{Prec} + \text{Rec}}.$$

*Calibration and resolution.*