
# Human-Centered Exploration of Table Unionability

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

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

Roee Shraga
Worcester Polytechnic Institute
Worcester, Massachusetts, USA
rshraga@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 Roee Shraga. Human-Centered Exploration of Table Unionability. PVLDB, 19(9): 2099 - 2112, 2026.
doi:10.14778/3819518.3819537

**PVLDB Artifact Availability:**
The source code, data, and/or other artifacts have been made available at https://github.com/NinaKlimenkova/TUNE_Benchmark.

## 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, 5, 7, 15, 21], 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, 11, 16, 18, 20, 22, 32, 35]. In this paper, we focus specifically on TUS and the challenges involved in defining, evaluating, and modeling unionability decisions.

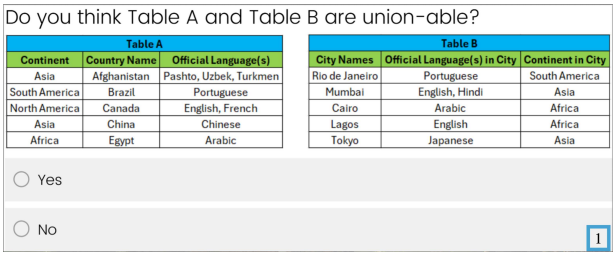


**Figure 1: Ambiguous unionability case from the survey.**

Despite the apparent simplicity of the task, the literature exhibits diverse interpretations of what "unionable" means [16, 20], and even domain experts often disagree when assessing the same table pairs [30]. Different frameworks adopt different criteria, from requiring identical schemas [37] to accepting domain-level correspondence [2, 32] to insisting on preserved inter-column relationships[20]. 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 [14, 20, 32, 34, 45]. In many cases, two domain experts may reasonably disagree on whether certain table pairs should be unioned, yet this disagreement is invisible in current evaluation frameworks. 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 [30], 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

*Both authors contributed equally to this research.


Proceedings of the VLDB Endowment, Vol. 19, No. 9 ISSN 2150-8097.
doi:10.14778/3819518.3819537

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. We provide a clear comparison in Section 2.4.

**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 unionability evaluation. TUNE's primary objective is to *assess human performance on the unionability decision task*, establishing a behavioral baseline for TUS by capturing votes together with confidence, decision time, interaction traces, and free-text explanations. From this human-centered foundation, we further study how richer human and algorithmic signals can *improve annotation quality and unionability prediction accuracy*. Finally, by releasing multi-annotator vote distributions, TUNE enables *quantifying pair-level ambiguity* and supports ambiguity-aware evaluation; we highlight this primarily as an opportunity enabled by TUNE for future work. These objectives support three intended use cases: evaluating TUS methods with richer human-alignment metrics beyond binary accuracy, developing human-in-the-loop data discovery pipelines informed by behavioral evidence and benchmarking LLM-based approaches against human baselines. With these use cases in mind, 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 [20], Starmie [16], 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[1] 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.

## 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.

[1] https://github.com/NinaKlimenkova/TUNE_Benchmark/tree/main/data

## 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* [16], *SANTOS* [20], 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. In our setting, a pair is labeled unionable when the two tables (generated on the same topic) contain a set of *unionable columns*, i.e., semantically similar columns that can be aligned across the tables. The *unionability decision problem* is to learn a function $f : \mathcal{T} \times \mathcal{T} \to [0, 1]$ 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 [37]. This definition ensures syntactic compatibility but proves overly restrictive for real-world data discovery, where tables from heterogeneous sources rarely share perfectly aligned schemas. Historically, schema matching can also be applied to identify corresponding attributes [23, 28, 36, 38, 39, 43].

**Relaxing Constraints: Domain-Based Union Search.** Nargesian et al. [32] 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), representing a more flexible, semantics-driven interpretation. Building on this, Bogatu et al. [2] introduced D3L, which constructs separate LSH indexes for five complementary evidence types—attribute name similarity, value overlap, format patterns, word embeddings, and domain distributions—and aggregates them through a learned weighted distance metric.

**Semantic Enrichment: Contextual Representations.** Fan et al. [16] advanced the notion of unionability by embedding tables into a semantic vector space, where the unionability score $f(T_q, T_c)$ emerges from geometric similarity in the learned embedding space rather than explicit domain matching. Subsequent work extends this direction: Hu et al. [18] propose *AutoTUS*, which jointly captures local column signals and global table structure, while Qiu et al. [35] introduce *LIFTus*, an adaptive multi-aspect column representation model for TUS that is designed to better handle non-linguistic and numeric data.

**Relationship-Aware Union Search.** Khatiwada et al. [20] introduced SANTOS, which 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* → *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:** These four families represent a clear progression from strict syntactic matching to increasingly flexible semantic and structural criteria, with each generation capturing more of what makes tables unionable. Yet all share fundamental limitations: they produce deterministic scores without modeling the uncertainty we observe among domain experts, contextual embeddings operate as black boxes with limited explainability, and relationship-aware definitions can be overly restrictive, excluding tables that humans nevertheless consider unionable (see Example 1). More broadly, none account for the context-dependent, interpretive nature of unionability that motivates the human-centered perspective we advance in this paper. Longer discussion is provided in our technical report [31].

## 2.3 Cognitive-aware Data Discovery

Cognitive aspects of human decision-making have been examined in related data integration tasks [24, 26, 48], most notably in schema and entity matching, where prior work [25, 40, 41] demonstrates that humans exhibit overconfidence, inconsistent evaluation strategies, and characteristic decision-time patterns. However, to the best of our knowledge, *no prior work has applied a similar cognitive, behavior-aware analysis to table unionability itself.* This gap is particularly important because unionability involves interpretation, context, and semantic reasoning that may vary across individuals, yet existing systems assume humans serve as reliable validators without examining the consistency or alignment of their judgments.

Complementary to this cognitive line of work, several recent papers critically examine TUS benchmarking. Pal et al. [33] propose a generative framework showing that LLM-generated benchmarks are substantially more challenging than hand-curated ones; Srinivas et al. [45] introduce LakeBench, demonstrating that current tabular foundation models still perform far from satisfactorily; and Boutaleb et al. [3] show that simple baselines can achieve surprisingly strong results due to dataset-specific artifacts. While these efforts call

for richer evaluations, they still treat unionability labels as fixed ground truth and do not analyze the human decision processes that generate those labels.

## 2.4 Preliminary Work

Our earlier workshop paper [30] took a first step toward addressing this gap through an exploratory study of human unionability judgments conducted over the same survey infrastructure described in Section 3.1. That study focused on initial behavioral observations and preliminary ML and LLM experiments.

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, indicating that humans struggle with certain unionability decisions and that this struggle is reflected in their interaction patterns. Training traditional ML models on these behavioral features achieved an average accuracy improvement of roughly 25 percentage points over raw human labels, with decision-time features alone providing much of the gain.

LLMs add another dimension to the study of unionability [17, 27, 29]. ALT-GEN [34] explored LLMs for TUS using natural-language table representations but did not incorporate human behavior. Our workshop study found that LLMs alone do not consistently outperform humans, yet when provided with human consensus and metacognitive metrics through in-context learning, LLM accuracy improved from 59.4% to 75.0%, matching 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 preliminary findings highlighted the need for a scalable, behavior-rich benchmark capturing how humans interpret unionability, motivating the development of TUNE. The current paper extends the workshop study by replacing its ad-hoc configurations with a systematic experimental framework involving 4 Scenarios and conducting a rigorous LLM evaluation across 8 models with selection based on calibration and resolution, as we described in Section 4. We also deepen the behavioral analysis through formal metacognitive metrics and sentiment-based explanation analysis, benchmark and integrate three TUS methods (Starmie[16], SANTOS[20], D3L[2]) as features, and formalize the resulting resource as the TUNE benchmark with explicit goals, use cases, and public data artifacts, as described in Section 3. Finally, we broaden annotator diversity through an additional annotation on Prolific[2], released as TUNE 2.0 that we described in Section 6.

# 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 by leveraging 26 table pairs from the UGEN benchmarks [33]. These pairs represent both unionable and non-unionable configurations and were strategically selected to ensure balanced difficulty across four survey versions (V1–V4). This design enables cross-version generalization analysis via leave-one-version-out evaluation while ensuring each participant sees a manageable number of questions. We inherit the binary reference labels from UGEN's ground-truth files[3]. In UGEN, a pair is labeled *unionable* when the two tables (generated on the same topic) contain a set of *unionable columns*—i.e., semantically similar columns that can be aligned across the tables; non-unionable pairs (often from the same topic) are constructed to have no such alignable columns [33]. While this criterion is consistent with the broader domain/semantic correspondence view of unionability [32], stricter definitions exist (e.g., schema identity [37] or relationship preservation [20]). We adopt UGEN's labeling criterion to maintain comparability.

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

| Statistic | Value |
|---|---|
| Table pairs | 26 |
| Survey versions | 4 (V1–V4) |
| Unionability questions (per version) | 8 |
| Unionability questions | 32 |
| Participants | 58 |
| Total judgments | 464 |
| Judgments per pair (avg.) | ≈ 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 |

We deployed a structured online survey through Qualtrics[4], 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[5].

[2]https://www.prolific.com

[3]https://github.com/northeastern-datalab/gen/blob/main/data

[4]https://wpi.qualtrics.com/jfe/form/SV_8jqmFQVl43NcHci

[5]https://github.com/NinaKlimenkova/TUNE_Benchmark

## 3.2 TUNE Goal and Use Cases

TUNE's primary objective is to *assess human performance on the unionability decision task*, establishing a behavioral baseline for TUS. It characterizes how accurately experts judge unionability and how confidence, decision time, interaction traces, and explanations relate to judgment reliability and reasoning strategies. Building on this human-centered foundation, TUNE supports a secondary objective of *improving annotation quality and unionability prediction accuracy* by combining richer human signals with algorithmic evidence, as we demonstrate in Scenarios S1–S4, Section 4. In addition, while the final unionability decision is binary, TUNE's multi-annotator vote distributions enable a richer, distributional view of ground truth that supports quantifying pair-level ambiguity. The proportion of annotators judging a pair as unionable serves as a soft unionability score, analogous to graded relevance in information retrieval. Just as the IR community evolved from binary relevance [8–10, 46] to graded scales evaluated with NDCG [19, 44], TUS evaluation can move toward ambiguity-aware metrics that distinguish near-unanimous consensus from contested decisions. We do not pursue this direction in the current paper, but we highlight it as a concrete opportunity that TUNE's design affords. These goals can potentially support three classes of systems. First, TUS method developers can use TUNE for evaluation beyond binary accuracy. Because each table pair carries multi-annotator votes and confidence scores, developers can diagnose whether a method's errors fall on genuinely ambiguous pairs or on clear consensus cases, using TUNE's human signals as an offline diagnostic lens. Second, designers of human-in-the-loop data discovery pipelines can use TUNE's behavioral metadata to inform routing and aggregation decisions. Such strategies have proven effective in related tasks like schema matching [40, 41] and crowdsourced entity resolution [24, 25], but do not yet exist for TUS, partly because no benchmark has provided the needed behavioral data. TUNE's decision-time, confidence, and click-pattern signals can support developing analogous policies, with our Scenarios S1–S4 (Section 4) serving as prototypes. Third, researchers benchmarking LLMs for data integration can use TUNE as a controlled testbed, comparing LLM predictions against both binary ground truth and human judgments, and exploring which behavioral signals are most effective as in-context evidence.

## 3.3 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. After explicitly cleaning the exported raw data file from Qualtrics, we initially had 21 attributes, capturing participant metadata, demographics, behavioral indicators (e.g., click counts, confidence, decision time, explanations), and question-specific attributes. This was later extended to 36 attributes through various types of normalization, label encoding, derived metrics, and feature engineering, including both human-annotated labels and ground truth labels from prior benchmarks to capture participant behavior and data discovery task complexity. 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.[6] 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%[30]. This systematic gap between expressed certainty and actual correctness reflects a stable overconfidence bias, a well-documented phenomenon in human decision-making[42]. By embedding these metacognitive signals directly into the benchmark, TUNE enables analyses that go beyond simple correctness labels. In Section 3.4, 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. Using a DistilBERT model fine-tuned on SST-2[7], we find that explanations skew strongly negative (285 vs. 102 positive), and that sentiment aligns with judgment direction: positive explanations are more often correct for "yes" judgments (75.4%), while negative

[6] https://github.com/NinaKlimenkova/TUNE_Benchmark/tree/main/data
[7] https://huggingface.co/distilbert/distilbert-base-uncased-finetuned-sst-2-english

explanations are more often correct for "no" judgments (72.1%). We also observe that explanations using relational terms ("related", "similar", "overlap") were more accurate (66.8%) than those relying on categorical notions such as "concept" or "type" (56.9%), suggesting that attention to structural relationships is a more reliable reasoning mode for this task. Further details are provided in our technical report [31].

**Post-survey questionnaire response.** Participants also reflected on what makes tables unionable by selecting union-defining criteria (e.g., common schema, same domain, same concept) for a small set of example pairs. They were generally accurate on clearly unionable examples (about 89%) but less so on non-unionable ones (about 62%), hinting that recognizing valid unions may be easier than ruling out borderline cases. Among the most frequently selected criteria were "common schema", "same domain", and "same concept", indicating that schema alignment and domain-level consistency are salient cues for respondents.

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.

## 3.4 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. Raw single-human accuracy ranges from 58–70%, while restricting 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 higher in all versions except V3, where it is essentially comparable (88% vs. 87.5%), likely due to sampling variability from restricting the vote to ten annotators. 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 [40], we evaluate two key dimensions of judgment quality: *calibration* and *resolution*.

Calibration measures how well participants' 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 [40], we define calibration for a group $g$ (e.g., a survey version) as $\mathrm{Cal}(g) = \bar{c}_g - \mathrm{Acc}_g$, where $\bar{c}_g$ is the mean reported confidence and $\mathrm{Acc}_g$ is the empirical accuracy. Values near zero indicate better calibration; positive values indicate overconfidence and negative values underconfidence.

**Table 2: Calibration and resolution metrics by response type.**

| Survey | Calibration | | Resolution | |
|---|---|---|---|---|
| Version | Yes | No | Yes | No |
| V1 | **0.140** | **0.110** | 0.037 | 0.196 |
| V2 | 0.205 | 0.238 | 0.018 | −0.278 |
| V3 | 0.142 | 0.118 | **0.391** | −0.026 |
| V4 | 0.178 | 0.194 | 0.079 | **0.478** |

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 [40], we measure resolution for a group $g$ using Goodman–Kruskal's rank correlation: $\mathrm{Res}(g) = \gamma(\mathbf{c}_g, \mathbf{y}_g)$, 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 2 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 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.

## 3.5 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 [16], SANTOS [20], and D3L [2]. These methods represent

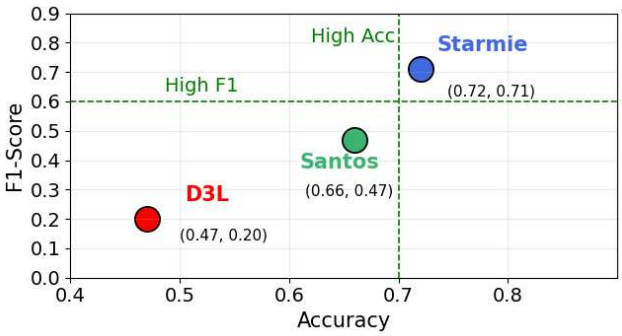


**Figure 2: TUS Methods Performance on TUNE. Points represent average performance across all survey versions.The x-axis denotes accuracy and the y-axis denotes F1-score.**

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 it's 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[8].

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 supervised learning could improve prediction beyond any single method. We trained standard ML classifiers (LR [13], KNN [12], RF [4], XGB [6]) using TUS scores as input features under a Leave-One-Version-Out (LOVO) cross-validation strategy (see Section 4), and conducted ablations over all feature combinations. As single-feature inputs, Starmie reaches 75% accuracy, SANTOS 66%, and D3L only 44%. Using all three together yields 75%, no better than Starmie alone, suggesting the model gains little from the weaker scores. The best configuration is SANTOS+Starmie at 78%, indicating that SANTOS contributes complementary signal while D3L tends to dilute performance. Detailed per-version and model-specific results are provided in our repository[9].

[8]https://github.com/NinaKlimenkova/TUNE_Benchmark/blob/main/Technical_Report.pdf

## 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.

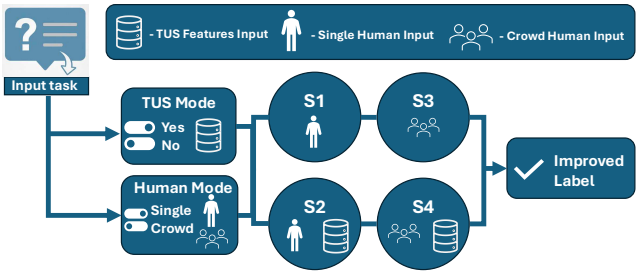


**Figure 3: Experimental unionability evaluation pipeline.**

### 4.1 Experimental Design Overview

Figure 3 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,

[9]https://github.com/NinaKlimenkova/TUNE_Benchmark/tree/main/results

we select the subset of features consistent with the active modes in Figure 3 (e.g., excluding TUS scores in S1) and train standard classifiers (LR, KNN, RF, and XGB) on this fixed representation. In all scenarios, ML models are trained *supervised* to predict the binary reference label $y$ for the underlying table pair. For the scenarios (S1/S2), the same $y$ is associated with each individual response to that pair, whereas for crowd-level scenarios (S3/S4) we train on one aggregated instance per question/table pair. For each scenario we run an ablation over feature groups and classifier types and then we interpret "which signals matter" via drop-group/only group ablations (performance changes when adding/removing feature groups); 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.[10]

**LLM Configuration: Model Selection and Prompting.** Despite growing interest in applying LLMs to data discovery tasks [27, 34], their ability to make unionability judgments, in particular to utilize rich human behavioral signals, remains largely unexplored. Prior work such as ALT-GEN [34] evaluates LLMs on TUS using natural-language table representations, but does not incorporate human behavioral metadata or study LLM behavior under controlled human-signal configurations. We include LLMs in our evaluation as a distinct modeling paradigm: rather than training a classifier on fixed features, we test whether in-context learning can serve as an alternative signal-integration mechanism, where human metadata and TUS scores are provided directly in the prompt. This allows us to compare how feature-based ML models and prompt-based LLMs behave under the same information configurations (S1–S4), relative to human performance, without task-specific model training.

For the LLM experiments, we first evaluated eight open-source models through the Groq API[11] on TUNE questions. For each model, we collected (i) a binary unionability decision and (ii) a self-reported confidence score on a 0–100 scale, similar to our participants. Based on accuracy and confidence reliability (calibration and resolution), we selected Qwen-3-32B [47] as our primary model. We then instantiated the same four information configurations (S1–S4) directly in prompt form, mirroring the switches in Figure 3. The detailed prompt templates and prompting procedure are provided in the technical report. Our goal is not to rank ML classifiers against the LLM, but to observe how feature-based ML and prompt-based LLMs behave 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 3.4 and are provided in detail in the technical report.

*Note on Scope:* Figures 5–11 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.

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 4), we report a representative configuration selected from a broader ablation over classifiers and feature groups, where *DecisionTime* emerged as the strongest single-annotator feature group on average (and produced the largest drop when removed): an XGB classifier trained only on the *DecisionTime* 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

[10] https://github.com/NinaKlimenkova/TUNE_Benchmark/tree/main/results
[11] https://console.groq.com/docs/overview

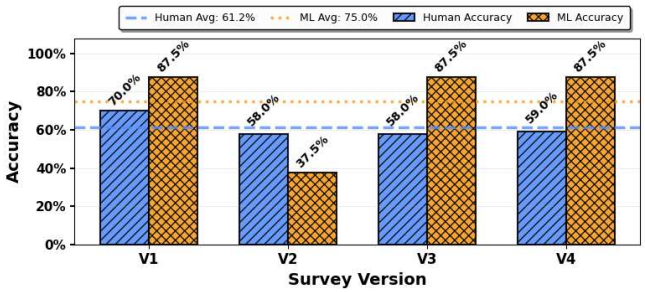


**Figure 4: ML Improvement Over Single Human Judgments.**

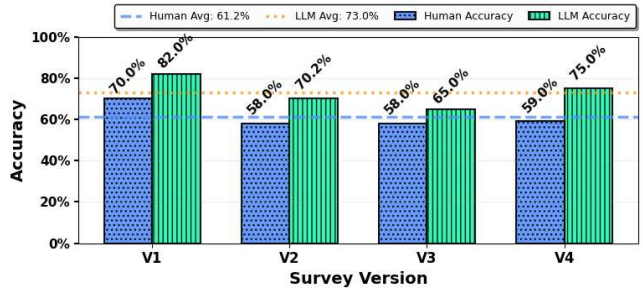


**Figure 5: LLM Evaluation of Single Human Judgments.**

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. Because the full ablation spans multiple classifiers and feature-group combinations across four LOVO folds, we summarize results in the main text by reporting per-version accuracy for a best representative configuration and describing the key ablation takeaway, while providing full per-version, per-feature, and per-model ablation tables in the technical report [31].

**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 3.4. At the version 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 metadata, 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 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 6 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 7), 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

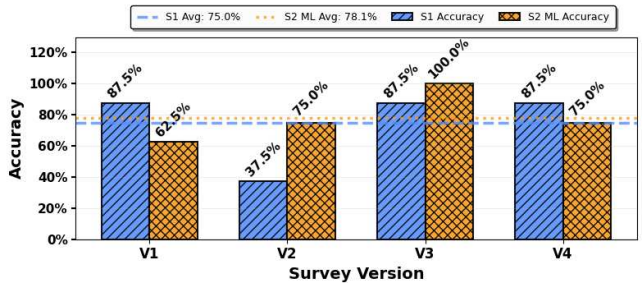


**Figure 6: Scenario 1 vs. 2 (ML): Impact of Adding TUS Features to Single Human Input.**

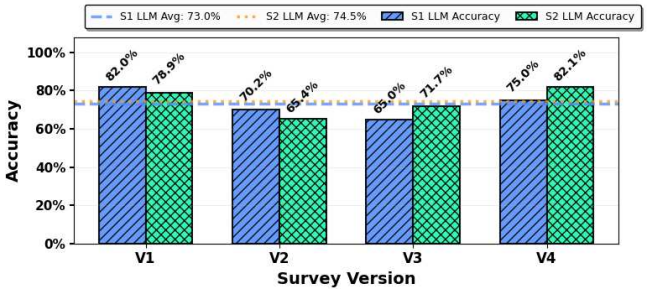


**Figure 7: Scenario 1 vs. 2 (LLM): Impact of Adding TUS Scores to Single Human Input**

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:**

- **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

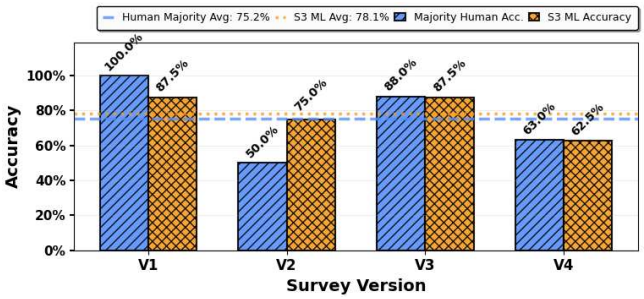


**Figure 8: ML Evaluation of Crowd Judgements.**

(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 8 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 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.

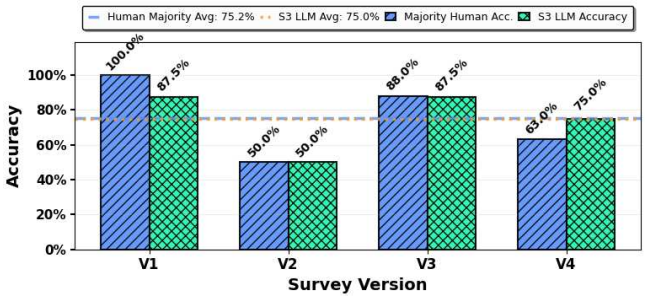


**Figure 9: LLM Evaluation of Crowd Judgements.**

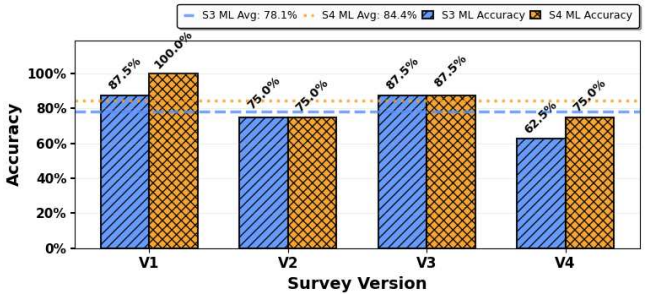


**Figure 10: S3 vs. S4 ML Accuracy across survey versions.**

- **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.

## 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:** Adding TUS scores to crowd-level features (Figure 10) yields the strongest performance across all scenarios: a KNN classifier using all feature groups reaches 84.4% average accuracy, up from 78.1% in S3, with 100% on V1 and gains on V4 (62.5% → 75.0%). Ablation results confirm that this gain is driven by the interaction between TUS scores and crowd vote statistics—removing either reduces accuracy to about 65.6%—while entropy and confidence aggregates prove dispensable.

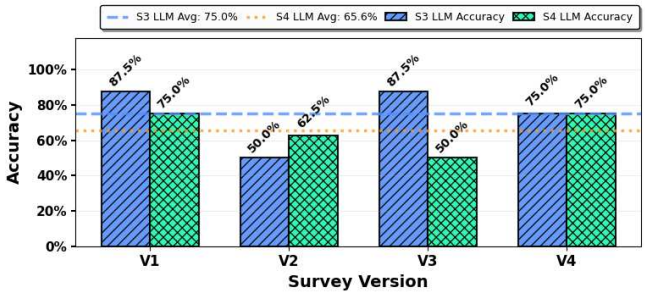


**Figure 11: S3 vs. S4 LLM Accuracy across survey versions.**

**LLM Results:** As shown in Figure 11, 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 for unionability. Individual judgments are informative but noisy. Overconfidence is common and behavioral patterns vary across versions, while majority voting, especially over stronger annotators, substantially improves reliability. The four scenarios illustrate how different information configurations improve labels: single-human behavioral traces serve as useful second opinions (S1), TUS scores dominate when available (S2), crowd aggregates form a strong baseline that ML and LLMs modestly refine (S3), and combining crowds with TUS yields the strongest ML performance at 84.4% (S4). However, LLM performance becomes more mixed in S4, showing that hybrids 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 that explicitly account for disagreement and ambiguity, leveraging distributional labels, multi-perspective annotations, and uncertainty-aware measures.

## 6 TUNE 2.0

To broaden the representativeness of TUNE, we conducted an additional annotation round using Prolific[12], recruiting 60 new participants with diverse professional backgrounds, education levels, and demographics beyond our original institutional setting. We release the extended dataset as TUNE 2.0. Key statistics are summarized in Table 3 alongside the original TUNE for comparison.

The TUNE 2.0 sample is substantially more diverse: participants span five age groups (78.4% between 25–54, vs. TUNE's 18–30 student cohort), represent fields including Computing/IT, Natural Sciences (26.7% each), Business/Finance (21.7%), and Engineering (13.3%), and are highly educated (60% bachelor's, 35% postgraduate) with high English proficiency (81.7% native speakers). Both individual and collective accuracy are higher in TUNE 2.0: single-human accuracy rises from 61.2% to 67.3% and majority-vote accuracy from 75.2% to 81.3%, while the approximately 14-point individual-to-majority gap remains stable, confirming that crowd aggregation benefits hold regardless of population composition. Calibration and resolution, computed following Section 3.4, reveal that overconfidence persists but is attenuated: calibration values range from −0.023 to +0.143 for "Yes" and +0.063 to +0.139 for "No" responses.

[12]https://www.prolific.com

**Table 3: Key statistics: TUNE vs. TUNE 2.0.**

| Statistic | TUNE | TUNE 2.0 |
|---|---|---|
| Table pairs | 26 | 26 |
| Survey versions | 4 (V1–V4) | 4 (V1–V4) |
| Unionability questions | 32 | 32 |
| Participants | 58 | 60 |
| Total judgments | 464 | 480 |
| Judgments per pair (avg.) | ≈14.5 | 15.0 |
| Accuracy (single human) | 61.2% | 67.3% |
| Accuracy (majority vote) | 75.2% | 81.3% |
| Confidence (per version) | 71–80% | 73.6–80.1% |
| Decision time (per question) | 85 s | 64.7 s |
| Click count (per question) | 2.2 | 2.04 |

Resolution remains modest overall, though V4 "Yes" yields a statistically significant correlation of 0.350 ($p < 0.05$), the only significant result across both datasets. The persistence of negative resolution for some "No" groups confirms that difficulty introspecting on rejection correctness is a stable cross-population phenomenon. TUNE 2.0 provides the community with a demographically and professionally heterogeneous counterpart to the original benchmark. This will enable researchers to study whether unionability patterns generalize across annotator populations, to develop and evaluate models on more representative data, and to investigate how annotator diversity affects judgment quality. We will consider these important directions for future work. The dataset, including raw responses, behavioral metadata, and computed features, is publicly available.[13]

## 7 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 four experimental scenarios, we show that 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 act as useful second opinions but are sensitive to conflicting signals.

These findings suggest that data discovery systems should favor hybrid architectures over purely human- or model-centric designs. By making behavioral indicators available alongside traditional labels, TUNE enables researchers to study not just *what* humans decide, but *how* and *why*. With TUNE 2.0, we further extend this foundation to a demographically diverse annotator pool, enabling cross-population generalization studies and supporting more representative, human-aligned unionability assessments.

## ACKNOWLEDGMENTS

This work was supported in part by NSF under award number IIS-2348121 and by a grant from the United States-Israel Binational Science Foundation (BSF), Jerusalem, Israel.

[13]https://github.com/NinaKlimenkova/TUNE_Benchmark/tree/main/data

# 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. In *Proceedings of the 4th Table Representation Learning Workshop*, Shuaichen Chang, Madelon Hulsebos, Qian Liu, Wenhu Chen, and Huan Sun (Eds.). Association for Computational Linguistics, Vienna, Austria, 71–85. https://doi.org/10.18653/v1/2025.trl-1.7

[4] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32. https://doi.org/10.1023/A:1010933404324

[5] 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

[6] 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

[7] 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 *Proceedings of the 28th International Conference on Extending Database Technology (EDBT)*. OpenProceedings.org, 397–410. https://doi.org/10.48786/edbt.2025.32

[8] Cyril Cleverdon and Michael Keen. 1966. *Factors Determining the Performance of Indexing Systems*. Cyril Cleverdon, Wharley End, Bedford, England. ASLIB Cranfield Research Project, Vol. 2.

[9] C. W. Cleverdon. 1960. *ASLIB Cranfield Research Project: Report on the First Stage of an Investigation into the Comparative Efficiency of Indexing Systems*. Technical Report. The College of Aeronautics, Cranfield, England. Aslib Cranfield Research Project.

[10] Cyril W. Cleverdon. 1962. *Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems*. Technical Report. Aslib Cranfield Research Project, Cranfield, England.

[11] Tianji Cong, Fatemeh Nargesian, and H. V. Jagadish. 2023. Pylon: Semantic Table Union Search in Data Lakes. *CoRR* abs/2301.04901 (2023). https://doi.org/10.48550/arXiv.2301.04901

[12] 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

[13] David R. Cox. 1958. The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society: Series B* 20, 2 (1958), 215–242.

[14] 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

[15] 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

[16] 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

[17] 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. https://par.nsf.gov/biblio/10656713

[18] 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

[19] Kalervo Järvelin and Jaana Kekäläinen. 2000. IR Evaluation Methods for Retrieving Highly Relevant Documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '00)*. ACM, Athens, Greece, 41–48. https://doi.org/10.1145/345508.345545

[20] Aamod Khatiwada, Grace Fan, Roee 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

[21] Aamod Khatiwada, Harsha Kokel, Ibrahim Abdelaziz, Subhajit Chaudhury, Julian Dolby, Oktie Hassanzadeh, Zhenhan Huang, Tejaswini Pedapati, Horst Samulowitz, and Kavitha Srinivas. 2024. TabSketchFM: Sketch-Based Tabular Representation Learning for Data Discovery Over Data Lakes. *2025 IEEE 41st International Conference on Data Engineering (ICDE)* (2024), 1523–1536. https://api.semanticscholar.org/CorpusID:270878553

[22] Aamod Khatiwada, Roee Shraga, and Renée J. Miller. 2026. Diverse Unionable Tuple Search: Novelty-Driven Discovery in Data Lakes. In *Proceedings of the 29th International Conference on Extending Database Technology (EDBT)*. OpenProceedings.org, 42–55. https://doi.org/10.48786/edbt.2026.04

[23] Christos Koutras, George Siachamis, Andra Ionescu, Kyriakos Psarakis, Jerry Brons, Marios Fragkoulis, Christoph Lofi, Angela Bonifati, and Asterios Katsifodimos. 2020. Valentine: Evaluating Matching Techniques for Dataset Discovery. *2021 IEEE 37th International Conference on Data Engineering (ICDE)* (2020), 468–479. https://api.semanticscholar.org/CorpusID:222378204

[24] Guoliang Li. 2017. Human-in-the-loop data integration. *Proc. VLDB Endow.* 10, 12 (Aug. 2017), 2006–2017. https://doi.org/10.14778/3137765.3137833

[25] 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

[26] Huanyu 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), e15. https://doi.org/10.1007/978-3-319-46523-4_13

[27] Kaiyu Li, Zhongxin Hu, Yuxin Gao, and Yuyang Wu. 2025. DeepSearch: LLM-powered Data Acquisition for Machine Learning. In *Proceedings of the 2nd International Workshop on Data-driven AI (DATAI), co-located with VLDB 2025*. https://www.vldb.org/2025/Workshops/VLDB-Workshops-2025/DATAI/DATAI25_11.pdf

[28] Yurong Liu, Eduardo H. M. Pena, Aécio Santos, Eden Wu, and Juliana Freire. 2025. Magneto: Combining Small and Large Language Models for Schema Matching. *Proc. VLDB Endow.* 18, 8 (April 2025), 2681–2694. https://doi.org/10.14778/3742728.3742757

[29] Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. 2025. DiscoveryBench: Towards Data-Driven Discovery with Large Language Models. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=vyflgpwfJW

[30] Sreeram Marimuthu, Nina Klimenkova, and Roee 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

[31] Sreeram Marimuthu, Nina Klimenkova, and Roee Shraga. 2026. Technical Report on TUNE. https://github.com/NinaKlimenkova/TUNE_Benchmark/blob/main/Technical_Report.pdf

[32] 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

[33] Koyena Pal, Aamod Khatiwada, Roee Shraga, and Renée Miller. 2023. Generative Benchmark Creation for Table Union Search. https://doi.org/10.48550/arXiv.2308.03883

[34] Koyena Pal, Aamod Khatiwada, Roee 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

[35] 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

[36] 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

[37] Raghu Ramakrishnan and Johannes Gehrke. 2003. *Database Management Systems* (3 ed.). McGraw–Hill.

[38] Nabeel Seedat and Mihaela van der Schaar. 2024. Matchmaker: Self-Improving Compositional LLM Programs for Table Schema Matching. In *3rd Table Representation Learning Workshop (TRL) at NeurIPS 2024*. https://openreview.net/forum?id=KCklcYUlLb

[39] Eitam Sheetrit, Menachem Brief, Moshik Mishaeli, and Oren Elisha. 2024. ReMatch: Retrieval Enhanced Schema Matching with LLMs. *CoRR* abs/2403.01567 (2024). https://doi.org/10.48550/arXiv.2403.01567

[40] Roee 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.

[41] Roee 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

[42] Roee 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

[43] Roee 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

[44] Eero Sormunen. 2002. Liberal Relevance Criteria of TREC: Counting on Negligible Documents?. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '02)* (Tampere, Finland). ACM, New York, NY, USA, 324–330. https://doi.org/10.1145/564376.564433

[45] 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

[46] Ellen M. Voorhees and Donna K. Harman. 2001. Overview of the Ninth Text REtrieval Conference (TREC-9). In *Proceedings of the Ninth Text REtrieval Conference (TREC-9) (NIST Special Publication)*. National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA. https://trec.nist.gov/pubs/trec9/papers/overview_9.pdf TREC-9 held Nov. 13–16, 2000; proceedings published Oct. 2001.

[47] 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

[48] Chen Jason Zhang, Lei Chen, H. V. Jagadish, Mengchen Zhang, and Yongxin Tong. 2020. Reducing Uncertainty of Schema Matching via Crowdsourcing with Accuracy Rates. *IEEE Transactions on Knowledge and Data Engineering* 32, 1 (2020), 135–151. https://doi.org/10.1109/TKDE.2018.2881185