How Should LLMs Assist Humans in Table Unionability Tasks?

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

Roee Shraga
Worcester Polytechnic Institute
Worcester, Massachusetts, USA
rshraga@wpi.edu

Erin Solovey
Worcester Polytechnic Institute
Worcester, Massachusetts, USA
esolovey@wpi.edu

ABSTRACT

Large language models (LLMs) increasingly take part in data discovery decisions such as table unionability—judging whether two tables can be meaningfully combined. Once framed as a model prediction or human-labeling problem, unionability is increasingly settled through LLM-assisted decision-making, where a person decides but a model's suggestion becomes one more input. The design question is then not only whether to provide assistance, but how to present it. We report a pilot survey in which participants judge a table pair, then may revise after seeing assistance in one of three forms: a bare recommendation, a recommendation with an explanation, or one that also conveys the model's expressed uncertainty. Across 90 decisions, aggregate accuracy barely moved, yet a transition-level view shows assistance corrected some errors while introducing a comparable number of new ones, and assistance form was associated with differences in both answer revision and confidence calibration. These preliminary findings suggest treating assistance form as a policy variable in future agentic data-discovery systems that expose, suppress, or qualify a suggestion according to the predicted risk of harmful reliance.

VLDB Workshop Reference Format:
Nina Klimenkova, Roee Shraga, and Erin Solovey. How Should LLMs Assist Humans in Table Unionability Tasks?.

VLDB Workshop Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://tinyurl.com/UnionHumanLLM.

1 INTRODUCTION

Large language models (LLMs) are moving from the edges of data discovery into its core, where they increasingly take part in semantic decisions, which were traditionally reserved for humans [8, 9, 15, 20]. One such decision is *table unionability*: judging whether two tables can be meaningfully combined [17]. It has commonly been framed in two main ways. A *model-prediction* problem, in which a learned model scores whether two tables describe the same kind of entity [7, 11, 17]. Alternatively, as a *human-labeling* problem, in which human judgments are used to curate ground-truth labels for training and evaluating such systems [13, 18]. In practice, however, the two often converge in LLM-assisted decision-making: a human makes the unionability call, but a language model's suggestion is placed in front of them and becomes one more input to their judgment [4]. In this workflow, the key question is not only whether LLM assistance is available but *how that assistance should be presented* to the human.

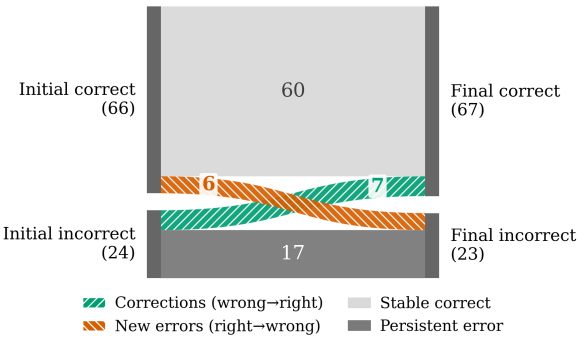

Figure 1: Transitions in participant answer correctness before and after LLM assistance.

Yet humans are imperfect at this task. An earlier study reports an accuracy of about 61% [13, 16], suggesting clear room for assistance. In many current settings, this assistance may come from an LLM. One might expect that placing an LLM suggestion in front of a human would straightforwardly improve the resulting judgments, but our pilot results suggest a more nuanced picture. Across 90 task decisions, aggregate accuracy changed very little after participants saw the LLM suggestion. Accuracy increased from 73.3% to 74.4%, which corresponds to only one additional correct answer. Taken alone, this near-flat result could suggest that assistance has little effect. Figure 1 tells a different story. Beneath the stable aggregate lie two opposing flows: assistance corrected 7 previously incorrect answers, but it also turned 6 previously correct answers wrong. The net gain of one masks roughly balanced help and harm. Assistance did not uniformly improve decisions so much as redistribute them, and final accuracy alone is blind to that redistribution.

This shifts the central question from whether assistance helps on average to when it helps, in what form, and under what conditions it may instead cause harm. To examine this, we designed and ran a survey in which each participant first judged a table pair and reported their confidence, then saw LLM assistance in one of three forms — (1) a recommendation-only, (2) a recommendation paired with a natural-language explanation, or (3) a recommendation with explanation accompanied by the model's expressed uncertainty. The participants could then revise both their answer and confidence. Recording behavior on both sides of the suggestion lets us observe revision and confidence change rather than only a final label. *The form of assistance appeared to matter:* it was

Proceedings of the VLDB Endowment. ISSN 2150-8097.

associated with whether answer changes tended to help or harm, and with how well the resulting confidence tracked correctness. Explanation-based assistance was the most promising in this pilot, while expressed uncertainty was associated with more harmful revisions and greater overconfidence.

The goal of this paper is to examine how the form of LLM assistance may influence human decision-making in table unionability. **Main contributions:**

- A preliminary transition-level analysis showing that a near-flat change in aggregate accuracy hides offsetting corrections and new errors.
- A comparison of three assistance forms—recommendation, explanation, and uncertainty—linking form to whether revisions help or harm and to confidence calibration.
- A vision for treating assistance form as a policy variable in future reliance-aware data-discovery systems.

2 LLM ASSISTANCE IN TABLE UNIONABILITY

This section positions our study within prior work on table unionability, human judgment in data integration and discovery, and LLM-assisted decision-making. We then describe the survey design, item selection, and measures used to examine how different forms of LLM assistance affect participants' answer revisions and confidence in table unionability decisions.

2.1 Related Work

Table unionability and data discovery. Table unionability is a core operation in data discovery, where the goal is to find tables that can extend a tabular dataset with additional rows [17, 21]. Most prior work treats it as an automated prediction task, scoring candidate tables by shared domains, column semantics, or learned representations [7, 11, 12, 17], with recent benchmarks using LLMs to generate and label union pairs at scale [19]. This has improved automated accuracy but largely abstracts away the human, treating unionability as a label to predict rather than a judgment a person makes. However, the judgment is often semantically ambiguous, with definitions emphasizing schema compatibility, domain overlap, semantic similarity, or the preservation of relationships among columns [13, 16]. This motivates studying not only model performance but how people reason about unionability.

Humans in the loop and the cognition of judgment. A complementary line keeps people in the loop for integration and matching, both as a source of labels and as decision-makers whose reliability must itself be understood [6, 14]. Studies of human matching behavior repeatedly document overconfidence and a gap between perceived and actual accuracy, and use behavioral signals such as decision time and confidence to characterize and calibrate judgment [1, 22, 24]. Our prior study brought this metacognitive lens to table unionability, finding confidence–accuracy gaps and showing that combining human and model signals can outperform either alone [13]; it did not, however, examine what happens to a person's judgment once an LLM suggestion is placed before them.

LLM-assisted decision-making and the form of assistance. A growing body of work studies how model predictions, explanations, and interface designs influence human decision-making [2, 10, 26]. The form of assistance matters: confidence scores can help calibrate trust yet do not by themselves guarantee better decisions [26]. Explanations can raise acceptance of a model's output, sometimes improving team performance, sometimes encouraging reliance even when the model is wrong [3, 5]. Recent LLM-based systems further explore human–LLM collaboration in search and information seeking through clarification, role-based prompting, retrieval augmentation, and user-guided disambiguation [4, 25].

What remains underexplored is how different forms of assistance compare specifically for table unionability, and how each affects not only whether a person revises an answer but whether that revision helps or harms and whether the resulting confidence stays calibrated. We take a first step toward this question, treating assistance form as a variable to be studied rather than fixed.

2.2 Study Design and Measures

Survey workflow. We study LLM-assisted unionability judgment through a survey-based prototype that separates a participant's independent judgment from their response to LLM assistance. Ten participants each answered nine table-pair questions, yielding 90 participant–task decisions. In each question, the participant first sees a pair of tables and gives an initial binary unionability decision (unionable/non-unionable), together with a confidence rating on a 0–100 slider. The survey then reveals LLM-generated assistance, after which the participant may revise their decision, update their confidence, and add a short free-text justification; they also rate how helpful the assistance was on a 1–5 scale. For every decision, we therefore record both pre-assistance and post-assistance answers and confidence ratings. Correctness is scored against the UGEN-derived benchmark label inherited from the prior unionability study [19]. Recording behavior on both sides of the LLM suggestion lets us observe answer revision and confidence change rather than only a final label. For more detailed view, we created an open version of our survey[1].

Assistance forms. Assistance is shown in one of three forms, with three questions assigned to each: a recommendation-only (the model's binary unionable/non-unionable prediction); a recommendation paired with a short natural-language explanation; or a recommendation and explanation accompanied by the model's self-reported uncertainty (a confidence percentage). The three forms are interleaved across the nine questions so that each participant encounters all of them within a single session.

Item selection. The nine table pairs were drawn from a larger candidate pool used in our prior unionability survey [16], whose table pairs and ground-truth unionability labels are themselves inherited from the UGEN benchmark [19]. We screened the pool so that the pilot would span a range of realistic conditions rather than only clear-cut cases, considering the difficulty and humans' performance of each question that was estimated from question-level accuracy in the earlier study [13]. LLM behavior was collected in advance using a single fixed prompt, recording the model's answer, its self-reported confidence, its explanation, and whether its answer aligned with the benchmark label. Guided by these signals, we selected a balanced set of items spanning easier and harder cases,

[1] https://wpi.qualtrics.com/jfe/form/SV_bPhtpgLuAatPCt0

Table 1: Accuracy and answer changes by assistance form.

Assistance form	Initial acc.	Final acc.	Switch rate	Beneficial changes	Harmful changes
Recommend. only	86.7%	86.7%	6.7%	1	1
Explanation	70.0%	83.3%	13.3%	4	0
Uncertainty	63.3%	53.3%	23.3%	2	5

unionable and non-unionable pairs, and both human–AI agreement and disagreement. More details about item selection strategy can be found in our repository[2].

Analysis measures. We characterize decisions along two axes: decision quality and confidence quality. For decision quality, we report accuracy before and after assistance and classify each paired decision into one of four transition outcomes: stable correct, corrected (incorrect → correct), new error (correct → incorrect), and persistent error. From these transitions, we derive a per-form switch rate, defined as the fraction of decisions whose answer changed, and counts of beneficial changes and harmful changes. This transition view is intended to distinguish movements that final accuracy alone would miss.

For confidence quality, we compute calibration and resolution following [23], before and after assistance, separately for each assistance form. Calibration captures how closely mean reported confidence $\bar{c}_g$ matches empirical accuracy Acc_g:

$$\mathrm{Cal}(g) = \bar{c}_g - \mathrm{Acc}_g. \tag{1}$$

With both quantities expressed on a common 0–100 scale, values near zero indicate well-aligned confidence, positive values suggest overconfidence, and negative values suggest underconfidence. Resolution captures whether confidence discriminates correct from incorrect decisions. We measure it as Goodman–Kruskal's rank correlation γ between the vector of confidence ratings $\mathbf{c}_g$ and the corresponding vector of correctness indicators $\mathbf{y}_g$:

$$\mathrm{Res}(g) = \gamma\left(\mathbf{c}_g, \mathbf{y}_g\right). \tag{2}$$

Larger positive values indicate that higher confidence tends to accompany correct decisions, values near zero indicate little discrimination, and negative values indicate that higher confidence is, on average, associated with errors. We also report the mean confidence shift from pre- to post-assistance to check whether an assistance form raises certainty without a corresponding improvement in correctness.

Together, these measures allow us to examine not only whether LLM assistance changed final accuracy, but also how participants moved between correct and incorrect answers and whether their confidence remained aligned with correctness.

3 PRELIMINARY FINDINGS

We now turn to preliminary findings, beginning with answer accuracy and reliance patterns before examining confidence quality. We read each result not only for what it reveals about participant behavior, but also for what it implies for a potential system.

[2]https://github.com/NinaKlimenkova/LLM-Assistance-for-Table-Unionability/blob/main/Questions_evaluation.xlsx

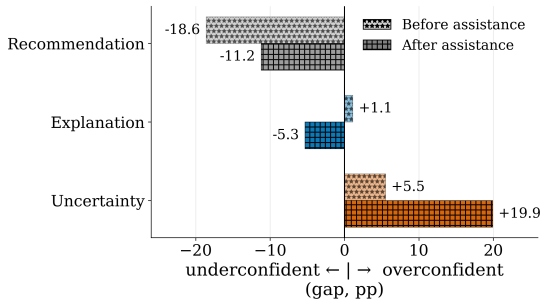

Figure 2: Calibration gap before and after assistance across assistance forms. Negative values indicate underconfidence, while positive values indicate overconfidence.

3.1 Decision Quality

We first examine whether LLM assistance improved participants' final unionability judgments. Across 90 participant–task decisions, initial accuracy was 73.3% (66/90) before participants saw the AI suggestion; after reviewing it, final accuracy rose only slightly to 74.4% (67/90). Taken alone, this suggests LLM assistance had a very modest effect on final task performance. But this small net change hides movement in the decision process. Figure 1 shows the transition from initial to final correctness. Most decisions were stable: 60 initially correct answers stayed correct and 17 initially incorrect remained incorrect. Of the 13 that changed, 7 were beneficial (incorrect → correct) and 6 harmful (correct → incorrect). The net gain of one thus masks two nearly balanced effects: assistance corrected some errors but introduced others.

Implication: Final accuracy alone collapses four distinct outcomes, stable correctness, corrected errors, persistent errors, and newly introduced mistakes, that carry different meaning for a system. The source of a final label matters. A correct answer may reflect the human's original judgment, a correction prompted by the AI, or an answer the AI simply confirmed. An incorrect answer may reflect either a human error that persisted or a new error introduced by the assistance. For future systems that reuse human feedback as labels for data integration, evaluation, or model refinement, these transitions should be logged rather than collapsed into a single accuracy score.

We next compare these patterns across the three assistance forms that we describe in Section 2.2. As shown in Table 1, recommendation-only assistance appeared mostly confirmatory: initial and final accuracy were both 86.7%, with a low switch rate of 6.7% and one beneficial and one harmful change. Explanation-based assistance showed the most promising pattern in this pilot, increasing accuracy from 70.0% to 83.3%, with four beneficial changes and no harmful changes. In contrast, the uncertainty condition produced the highest switch rate, at 23.3%, and more harmful than beneficial changes. These differences suggest that assistance form may shape not only whether users change their answers, but also whether those changes improve or degrade final decision quality. Because each assistance form is tied to specific table-pair instances in this pilot, we interpret these patterns as exploratory rather than causal.

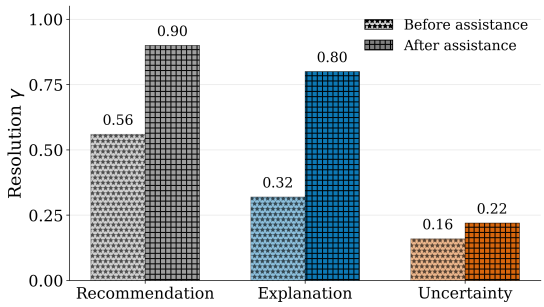

Figure 3: Resolution before and after assistance across assistance forms. Higher values indicate a stronger association between participants' confidence and correctness.

Implication: These patterns imply that assistance form could become a policy variable in future agentic data-management systems. Rather than always exposing the same LLM output, a system could adapt the form of assistance to the situation: recommendation-only support for lightweight confirmation, explanations for semantically ambiguous cases, and confidence-augmented assistance for cases that require more careful review. This points toward adaptive assistance mechanisms that balance decision support with the risk of over-influencing the human reviewer.

3.2 Confidence Quality

We next examine how assistance affected participants' confidence, and whether it remained informative of correctness, using the calibration and resolution measures from Section 2.2. Across all 90 decisions, mean confidence rose from 69.3 to 75.6 while accuracy barely moved (73.3% → 74.4%), suggesting assistance raised certainty more than correctness and making calibration and resolution central to assessing that confidence.

At the aggregate level, confidence became better aligned with correctness: the calibration gap moved from −4.0 (slight underconfidence) to +1.1 percentage points (near zero), and resolution rose from $\gamma = 0.27$ to $\gamma = 0.52$. Participants thus became both more confident and more informative of correctness.

These aggregate trends, however, varied by form. Figure 2 shows the calibration gap by form: recommendation-only assistance reduced underconfidence (−18.6 → −11.2) and explanation stayed near optimal (+1.1 → −5.3), while uncertainty showed the largest risk, widening into stronger overconfidence (+5.5 → +19.9).

Resolution followed a related but distinct pattern, shown in Figure 3. Recommendation-only sharply increased it (0.56 → 0.90) despite flat accuracy we see in Table 1, and explanation did too (0.32 → 0.80)—and given explanation's accuracy gains, that rise reflects genuinely more informative confidence. Uncertainty rose only slightly (0.16 → 0.22), raising certainty without making it better at separating correct from incorrect answers.

Implication: These patterns suggest that future agentic data-discovery systems should treat confidence not as a user-reported number but as a signal whose quality depends on the form of assistance: some forms make confidence more grounded and informative, while others raise certainty without improving its usefulness. This points toward assistance policies that monitor not only whether confidence rises but whether it stays calibrated and diagnostic enough to guide downstream actions such as accepting a label, requesting another review, or escalating an ambiguous pair.

Overall, these preliminary findings suggest that LLM assistance shaped decisions in ways final accuracy alone does not capture. Explanation-based support appeared most promising in this pilot—improved accuracy, beneficial revisions, and more informative confidence, without the overconfidence seen under uncertainty. This observation is pointing toward systems that adapt not only *whether* assistance is provided but *what form* it takes. We next discuss limitations and directions for future work.

3.3 Limitations and Future Work

As a pilot, this study is intended to surface hypotheses about LLM assistance rather than to settle them, and its scope is correspondingly modest: ten participants and 90 decisions, 30 per form. By design, each assistance form was paired with a fixed set of table pairs, so form and item are intertwined in these results. We therefore read the form-level patterns as exploratory—indications of where assistance form may matter, to be confirmed under more controlled conditions.

A larger study should counterbalance or randomize assistance forms across table pairs to separate form from item, add participants, and test multiple LLMs and prompting strategies, allowing firmer conclusions about when explanations help, when a bare recommendation suffices, and when expressed uncertainty increases rather than reduces harmful reliance. It should also conduct deeper analysis of behavioral signals such as justifications, helpfulness ratings, decision time, and trajectories of confidence change, as candidate input to an adaptive policy.

Building such a policy is itself an open problem. A system would need to pick a form of assistance from signals available before a decision is final—how confident the human is, how ambiguous the item looks, and whether the human and model agree. It would then be judged by its effect: the errors it helps correct against the new errors it causes. A natural next step is to test such policies in data-discovery workflows, where human labels feed downstream tasks. The question is whether adapting the form of assistance improves not just the labels, but the reliability of the process behind them.

4 CONCLUSION

We reframed table unionability as a problem of LLM-assisted human decision-making, where the design question is not only *whether* to offer a suggestion but *how* to present it. Our pilot's central lesson is methodological. Overall accuracy barely changed, but that flat number hid two opposing movements: assistance corrected some answers but introduced new errors in others. The form of the assistance also mattered—it shaped whether changes helped or hurt, and whether people's confidence matched how often they were right. How a decision was reached, and how the suggestion was shown, are part of the outcome. If a system can adjust the form of assistance instead of fixing it in advance, it could help people decide without quietly overriding their judgment. Confirming when, and for whom, each form helps is the work ahead.

REFERENCES

[1] Rakefet Ackerman, Avigdor Gal, Tomer Sagi, and Roee Shraga. 2019. A Cognitive Model of Human Bias in Matching. In *PRICAI 2019: Trends in Artificial Intelligence: 16th Pacific Rim International Conference on Artificial Intelligence, Cuvu, Yanuca Island, Fiji, August 26–30, 2019, Proceedings, Part I* (Cuvu, Yanuka Island, Fiji). Springer-Verlag, Berlin, Heidelberg, 632–646. https://doi.org/10.1007/978-3-030-29908-8_50

[2] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S. Weld. 2021. Is the Most Accurate AI the Best Teammate? Optimizing AI for Teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 11405–11414. https://doi.org/10.1609/aaai.v35i13.17359

[3] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 81, 16 pages. https://doi.org/10.1145/3411764.3445717

[4] Gloris Denisse Cedeño Batista, Joemon Jose, and Hideo Joho. 2026. CollabSearch: A Study of User-LLM Collaboration in Task-Based Search. In *Proceedings of the 2026 Conference on Human Information Interaction and Retrieval (CHIIR '26)*. Association for Computing Machinery, New York, NY, USA, 254–263. https://doi.org/10.1145/3786304.3788848

[5] Valerie Chen, Q. Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. 2023. Understanding the Role of Human Intuition on Reliance in Human-AI Decision-Making with Explanations. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 370 (Oct. 2023), 32 pages. https://doi.org/10.1145/3610219

[6] AnHai Doan, Adel Ardalan, Jeffrey Ballard, Sanjib Das, Yash Govind, Pradap Konda, Han Li, Sidharth Mudgal, Erik Paulson, G. C. Paul Suganthan, and Haojun Zhang. 2017. Human-in-the-Loop Challenges for Entity Matching: A Midterm Report. In *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics* (Chicago, IL, USA) *(HILDA '17)*. Association for Computing Machinery, New York, NY, USA, Article 12, 6 pages. https://doi.org/10.1145/3077257.3077268

[7] 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

[8] Raul Castro Fernandez, Aaron J. Elmore, Michael J. Franklin, Sanjay Krishnan, and Chenhao Tan. 2023. How Large Language Models Will Disrupt Data Management. *Proc. VLDB Endow.* 16, 11 (July 2023), 3302–3309. https://doi.org/10.14778/3611479.3611527

[9] Juliana Freire, Grace Fan, Ben Feuer, Christos Koutras, Yurong Liu, Eduardo Peña, Aécio S. R. Santos, Cláudio Silva, and Eden Wu. 2025. Large Language Models for Data Discovery and Integration: Challenges and Opportunities. *IEEE Data Eng. Bull.* 49 (2025), 3–31. https://api.semanticscholar.org/CorpusID:277195386

[10] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) *(CHI '99)*. Association for Computing Machinery, New York, NY, USA, 159–166. https://doi.org/10.1145/302979.303030

[11] 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

[12] 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

[13] Nina Klimenkova, Sreeram Marimuthu, and Roee Shraga. 2026. Human-Centered Exploration of Table Unionability. *Proceedings of the VLDB Endowment* 19, 9 (2026), 2099–2112. https://doi.org/10.14778/3819518.3819537

[14] 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

[15] 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

[16] 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

[17] 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

[18] 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

[19] 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

[20] Marcel Parciak, Brecht Vandevoort, Frank Neven, Liesbet M. Peeters, and Stijn Vansummeren. 2025. LLM-Matcher: A Name-Based Schema Matching Tool using Large Language Models. In *Companion of the 2025 International Conference on Management of Data* (Berlin, Germany) *(SIGMOD/PODS '25)*. Association for Computing Machinery, New York, NY, USA, 203–206. https://doi.org/10.1145/3722212.3725112

[21] Norman W. Paton, Jiaoyan Chen, and Zhenyu Wu. 2023. Dataset Discovery and Exploration: A Survey. *ACM Comput. Surv.* 56, 4, Article 102 (Nov. 2023), 37 pages. https://doi.org/10.1145/3626521

[22] Roee Shraga. 2022. HumanAL: calibrating human matching beyond a single task. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics* (Philadelphia, Pennsylvania) *(HILDA '22)*. Association for Computing Machinery, New York, NY, USA, Article 7, 8 pages. https://doi.org/10.1145/3546930.3547496

[23] 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.

[24] 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

[25] Ryen W. White. 2024. Advancing the Search Frontier with AI Agents. *Commun. ACM* 67, 9 (Aug. 2024), 54–65. https://doi.org/10.1145/3655615

[26] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3351095.3372852