Strong Matchers, Weak Architecture: Towards Confidence Guided Global Consistency in LLM-based Record Linkage

Luma C. R. Pires[1], Jorge R. A. Junior[1]

[1]Departamento de Engenharia de Computação e Sistemas Digitais (PCS) – Escola Politécnica – Universidade de São Paulo (USP), São Paulo, Brazil

{lumacicilia,jorgerady}@usp.br

Abstract. *Record Linkage produces transitive inconsistencies that existing LLM-based methods address locally or by construction, but none optimizes at partition level. We investigate whether LLM confidence scores carry sufficient signal to drive global partition optimization in clean-clean multi-source settings. On MusicBrainz, scores discriminate true from false matches among predicted matches (mean 0.936 vs. 0.729). A global ILP exploiting scores as edge weights resolves all 42 inconsistent components by removing 52 edges (98.1% true false positives) - fewer than greedy alternatives and with an optimality guarantee. These results motivate LLM confidence scores as input to partition-level optimization, with efficient approximations as future work.*

1. Introduction

Record Linkage (RL) identifies records from different sources referring to the same real-world entity [BINETTE; STEORTS, 2022]. The fifth generation of RL leverages generative Large Language Models (LLMs) as zero-shot matchers [NIKOLETOS; IOANNOU; PAPADAKIS, 2024], eliminating the labeling bottleneck and achieving strong pairwise performance [PEETERS; BIZER, 2023; HUANG et al., 2025].

Despite these advances, a structural gap persists. The literature has increasingly recognized that independent pairwise decisions can produce globally inconsistent results: a single false positive suffices to incorrectly merge distinct entity clusters. Recent LLM-based works respond either by repairing clusters post-hoc [PARDO et al., 2025a, 2025b; CHRISTEN et al., 2025] or by building consistency into the matching step itself [FU et al., 2025; WANG et al., 2025], yet all share the same limitation: decisions are taken locally without a partition-level objective. The classical RL literature already proposed this path - Correlation Clustering (CC) with weighted edges encoding match confidence [BANSAL; BLUM; CHAWLA, 2002; DEMAINE et al., 2006], evaluated for multi-source settings [SAEEDI; PEUKERT; RAHM, 2017] using classical similarity scores - predating LLM-based matchers and their confidence signals - but it remains unexplored in the LLM context: existing work extracts confidence scores only for active learning [OKAYAMA; ITO; MORISHIMA, 2026] or ensemble voting [ZEAKIS et al., 2025], never used as continuous weights for global clustering.

This paper makes three contributions. First, we provide empirical evidence that GPT-4o-mini confidence scores carry discriminative signal: mean score is 0.931 for true matches versus 0.008 for non-matches across all 30,000 evaluated pairs (Mann-Whitney $p \approx 0$) and remains separated among predicted matches alone (0.936 vs. 0.729), confirming the signal is sufficient for partition optimization. Second, we show that

standard CC requires an injectivity constraint - each record may match at most one record per source - to correctly model the clean-clean multi-source RL setting: without it, 38 of 42 inconsistent components remain unresolved; with it, all are resolved. Third, we demonstrate that a global Integer Linear Programming (ILP) solver exploiting this signal eliminates all 42 detected inconsistencies with a 98.1% FP-rate on removed edges, outperforming greedy baselines that either ignore scores (64.8% FP-rate on average) or lack an optimality guarantee - suggesting LLM confidence scores as viable weights.

2. Background and Related Work

2.1. Transitivity in Entity Resolution

Given records from K sources, RL aims to find a partition where each cluster contains records referring to the same real-world entity. A correct partition must satisfy transitivity - if $r_i = r_j$ and $r_j = r_k$, then $r_i = r_k$ - and, in clean-clean settings, source consistency: each cluster contains at most one record per source [PARDO et al., 2025a; WANG et al., 2025]. Pairwise methods decompose this into independent binary decisions, which are structurally insufficient: treating record pairs as independent can produce logically impossible linkage configurations, and combining local decisions via transitive closure compounds these errors into large inconsistent clusters [BINETTE; STEORTS, 2022; DRAISBACH; CHRISTEN; NAUMANN, 2019].

Global graph partitioning offers a principled alternative. Correlation Clustering [BANSAL; BLUM; CHAWLA, 2002] finds the partition that maximizes agreement across all edges of a similarity graph, enabling any local decision to be revised under global evidence. Demaine et al. (2006) extend this to weighted edges encoding continuous match confidence. Although CC is NP-hard in general, efficient approximations have been applied to RL [BINETTE; STEORTS, 2022]. This formulation naturally enforces transitivity at partition level - yet it remains unexplored in the LLM-based RL context, where confidence scores could serve as principled edge weights.

2.2. The Pairwise Paradigm in Language Model-based RL

Language model adoption for RL progressed from encoder-based pairwise classifiers and embedding extractors [LI et al., 2020; PEETERS; BIZER, 2021; CAO et al., 2024] to generative LLMs enabling zero-shot matching via prompting [PEETERS; BIZER, 2023], eliminating the labeling bottleneck - though no single model dominates across scenarios [HUANG et al., 2025]. These methods evaluate each pair independently, offering no guarantee of transitivity across the full set of matching decisions.

Within the LLM-based RL literature, this limitation has gained explicit recognition: Okayama, Ito and Morishima (2026) and Wang et al. (2025) acknowledge that pairwise decisions can produce transitively inconsistent results, and Carvalho et al. (2025) show violations persist even when pairwise accuracy is high. Despite this recognition, existing responses remain local: none formulates the problem as a partition-level optimization.

2.3. Addressing Transitivity: Progress and Remaining Gaps

Consistency by construction. Wang et al. (2025) and Zeakis et al. (2025) enforce mutual exclusivity via selection prompts, but process anchors in independent calls without shared

memory, leaving source consistency unresolved globally: two anchors may select the same candidate. Fu et al. (2025) cluster record sets in-context, guaranteeing transitivity within each batch via a representative-based merging strategy; however, decisions across batches are irreversible, and early misclassifications or blocking errors can permanently merge distinct entity clusters. In all approaches, committed decisions cannot be revised under contradictory global evidence.

Post-hoc cluster repair. Pardo et al. (2025a) apply minimum edge cut and betweenness centrality to oversized components; Pardo et al. (2025b) iterate between inconsistency detection, LLM querying, and matcher fine-tuning, resorting to blind pruning when the labeling budget is exhausted; Cao et al. (2024) combine active learning with minimum cut; Christen et al. (2025) train a Random Forest on topological features using LLM-generated labels. All four share the same structural limitations: repair is local, each component processed independently without a partition-level objective; decisions rely exclusively on graph topology; and results are strongly dependent on the quality of the initial graph, as errors introduced by the pairwise matcher propagate into clusters that topology-based repair cannot reliably resolve [PARDO et al., 2025a, 2025b].

2.4. Confidence Scores as Signal

Okayama, Ito and Morishima (2026) extract match probabilities from a fine-tuned encoder model - the LLM serves only as a binary labeling oracle for active learning - and use transitive inconsistencies as a sampling signal, but the final model remains pairwise with no architectural guarantee against future global violations. Zeakis et al. (2025) use generative LLMs directly as matchers, eliciting qualitative confidence levels that are converted into numeric weights for ensemble voting; scores improve weaker models but are consumed within the matching step - decisions remain pairwise and independent, with no subsequent mechanism to check or repair global consistency. Li et al. (2024) use classical similarity functions to generate candidate partitions and submit selected record pairs to a generative LLM as binary yes/no questions, updating a probability distribution over partitions via Bayes' theorem until the budget is exhausted; however, the LLM contributes only binary responses rather than continuous confidence scores, the partition space is fixed at initialization, and validation is restricted to K=2 datasets. Cao et al. (2024) generate similarity scores via SBERT embeddings - without a generative LLM - restricting score usage to canonical record selection rather than conflict resolution.

Across all surveyed approaches, no method formulates transitivity enforcement as a global optimization over the full partition, and none routes LLM confidence scores as continuous edge weights into such an objective - the gap this paper addresses.

3. Empirical Evidence

3.1. Experimental Setup

We evaluate whether GPT-4o-mini confidence scores carry a useful signal for detecting and resolving transitive inconsistencies in multi-source RL. Experiments use MusicBrainz, a standard five-source dataset of music track records. We focus on three sources (K = 3), comprising 3,827, 3,933, and 3,926 records respectively, yielding three

pairwise source combinations. Candidate pairs are produced by the GraphCR blocking method [CHRISTEN et al., 2025], generating 115,800 pairs with 5,222 true matches.

To control API cost, we draw a stratified random sample of 30,000 pairs - 10,000 per source pair - preserving the natural match ratio. Each pair is submitted to GPT-4o-mini via a structured prompt providing title, artist, album, duration, and year for both records, instructing the model to return a JSON object with a single field *prob_match* ∈ [0,1]. Missing fields are treated as absent rather than contradictory; the prompt also instructs to normalize heterogeneous formats before comparing. A pair is predicted as a match when *prob_match* ≥ 0.5, prioritizing recall over precision - the resulting false positives are precisely the target of subsequent global resolution - and G is built with records as nodes and predicted matches as weighted edges.

We compare four conditions. C1 (Pairwise LLM) is the unmodified LLM output, the baseline. C2 (Greedy-Score) iteratively removes the lowest-scoring edge from each inconsistent component; ties are broken randomly. C3 (Greedy-Betweenness) removes the edge with highest betweenness centrality, a purely structural criterion that ignores scores; ties are broken randomly. For nodes *u*, *v*, *w* of a component graph *G*, with *s(·)* the source function and s_{uv} the edge score, C4 (ILP-CC + Injectivity) solves per inconsistent component: $max \sum_{(u,v)} (s_{uv} - 0.5)\, x_{uv}$, $x_{uv} \in \{0, 1\}$, indicating kept edges, subject to i) transitivity, $x_{uv} + x_{vw} - x_{uw} \leq 1$ (for all permutations of *u*, *v*, *w*) and ii) injectivity,

$$\sum_{b:\, s(b)=\sigma} x_{ab} \leq 1, \quad \forall a \forall \sigma \neq s(a),$$

solved exactly via CBC (PuLP).

All methods operate exclusively on the 42 inconsistent components; the remaining 1,116 consistent components are left unchanged, as they already satisfy all domain constraints. Although GPT-4o-mini returns values discretized to one decimal place (0.5, 0.6, ..., 1.0), scores are used as ordinal edge weights in C4 - their discriminative power is confirmed in Section 3.2.

3.2. Results and Discussion

Table 1. Score distribution by range on the 30,000-pair stratified sample

Score Range	Pairs	Matches	Precision
[0.0 – 0.1)	28,002	0	0.000
[0.1 – 0.5)	551	0	0.000
[0.5 – 0.7)	23	1	0.043
[0.7 – 0.8)	70	2	0.029
[0.8 – 0.9)	67	33	0.493
[0.9 – 1.0]	1,287	1,271	0.988

Score signal and discrimination. Although GPT-4o-mini returns values discretized to one decimal place (0.5, 0.6, ..., 1.0), the scores strongly discriminate true from false

matches: pairs scoring ≤ 0.7 are false positives in over 99.9% of cases (n = 28,646), pairs at 0.8 split nearly evenly between matches and non-matches (49.3% TPs, n = 67), and pairs scoring above 0.9 are true positives in 99.8% of cases (n = 463), while pairs at exactly 0.9 are true positives in 98.2% of cases (n = 824) - justifying their use as ordinal edge weights in the correlation clustering objective. Precision rises sharply from near zero below 0.8 to 0.988 above 0.9 (Table 1). Scores are strongly bimodal: 97.6% of non-matches receive a score below 0.1, while 99.8% of true matches score at or above 0.8. Among all 30,000 pairs, mean score is 0.931 for true matches vs. 0.008 for non-matches ($\Delta = 0.924$; Mann-Whitney U, $p \approx 0$). Restricting to predicted matches (score ≥ 0.5), mean score is 0.936 for true positives vs. 0.729 for false positives. Pairwise F1 at threshold 0.5 is 0.949 (Precision = 0.903, Recall = 1.000 within the evaluated sample).

Inconsistent components. Applying threshold 0.5 yields 1,158 connected components, of which 42 (3.6%) are inconsistent - i.e., contain two or more records from the same source, collectively involving 70 false-positive edges. A component is considered resolved when no two records from the same source remain in the same cluster, i.e., when it no longer violates injectivity. Of 333 potential transitivity triangles, 78.7% have an unevaluated third edge - pairs that were never compared. Inferring these via transitive closure without a confidence signal risks propagating false positives; this coverage gap is partly a consequence of sampling 30,000 of 115,800 candidate pairs, and a score-weighted global solver can handle missing edges without this risk.

Table 2. Comparison of resolution methods. FP-rate: fraction of removed edges that were true false positives

Method	Precision	Recall	F1	Incons.	Removed Edges	FP-rate
C1	0.903	1.000	0.949	42	-	-
C2	0.942 ± 0.001	0.999 ± 0.001	0.970 ± 0.001	0	60.4 ± 1.1	0.984 ± 0.012
C3	0.924 ± 0.002	0.986 ± 0.002	0.954 ± 0.002	0	53.3 ± 0.9	0.648± 0.054
C4	0.936	0.999	0.967	0	52	0.981

Resolution methods. Table 2 compares the four conditions in the inconsistent subgraph. C1 and C4 are deterministic. C2 and C3 report mean over 100 runs with random tie-breaking. C4 (ILP-CC + Injectivity) eliminates all 42 inconsistencies, removing 52 edges of which 51 (98.1%) are true false positives - only one true match is discarded. This one error occurs in a component where a false positive and a true positive share an identical score (0.80), leaving the solver without a discriminating signal. By contrast, C3 (Betweenness), which ignores LLM scores entirely, achieves 64.8% FP-rate on average (std = 5.4%, 100 runs): it removes on average 18.4 true matches. C2 (Greedy-Score) reaches 98.4% FP-rate on average - comparable to C4's 98.1% - but performs on average 60.4 removals - eight more than C4. The difference is structural: C2 processes one component at a time and removes one edge per iteration; each removal can split a component into two sub-components that are themselves inconsistent, triggering further removals in cascade (60 total for 42 initial components). C4 solves each component as a single global optimization, avoiding cascade entirely and offering an optimality guarantee

within each component. The injectivity constraint is the critical addition over standard Correlation Clustering: without it, 38 of 42 components remain unresolved because the solver has no incentive to separate nodes of the same source. The threshold $\tau = 0.5$ is not optimized; a higher value would reduce false positives at the cost of recall, and the interaction between threshold choice and global resolution quality remains an open question for future work.

4. Conclusion

Pairwise LLM-based Record Linkage does not guarantee globally consistent partitions - a limitation that becomes critical in multi-source or clean-clean settings. Our experiments demonstrate that GPT-4o-mini confidence scores carry discriminative signal within inconsistent components, and that Correlation Clustering with an injectivity constraint - the domain-specific adaptation for clean-clean RL - exploits this signal to resolve all 42 detected inconsistencies while discarding only one true match. Greedy baselines either ignore scores (low precision) or process components incrementally, triggering cascading removals that a single global optimization avoids - evidence that partition-level optimization, not just score-awareness, drives the result. These results are exploratory and based on a single dataset and model; generalization to other domains requires further investigation.

Four limitations bound these results: i) the experiment covers 30,000 of 115,800 candidate pairs; detection targets injectivity violations only, leaving most transitivity triangles unobserved (Section 3.2); ii) LLM scores are ordinal rather than calibrated probabilities, though their discriminative power suffices for the solver; iii) the ILP is exact but intractable at scale; and iv) querying the LLM for all candidate pairs is cost-prohibitive at scale. The central open problem is finding an efficient approximation of CC with injectivity that preserves the score signal, potentially combining cheap similarity proxies for clear-cut pairs with LLM queries reserved for ambiguous cases - reducing both computational and monetary cost without sacrificing partition-level optimality.

Acknowledgments

LLM assistance (Claude Sonnet 4.6, Anthropic) was used for English language editing and revision of text originally drafted in Portuguese.

References

Bansal, N.; Blum, A.; Chawla, S. Correlation clustering. In: **IEEE SYMPOSIUM ON FOUNDATIONS OF COMPUTER SCIENCE**, 43., 2002. Proceedings [...]. [S.l.: s.n.], 2002. p. 238–247.

Binette, O.; Steorts, R. (Almost) All of Entity Resolution. *Science Advances*, v. 8, 2022.

Cao, H. et al. Graph Deep Active Learning Framework for Data Deduplication. *Big Data Mining and Analytics*, v. 7, n. 3, p. 753–764, 2024.

Carvalho, G. H. de et al. Ice-ID: A Novel Historical Census Data Benchmark Comparing NARS against LLMs & A ML Ensemble on Longitudinal Identity Resolution. *arXiv preprint*, 2025.

Christen, V. et al. Graph Metrics-driven Record Cluster Repair meets LLM-based active learning. *Journal of Data and Information Quality*, v. 17, n. 2, 2025.

Demaine, E. D. et al. Correlation Clustering in general weighted graphs. *Theoretical Computer Science*, v. 361, n. 2, p. 172–187, 2006.

Draisbach, U.; Christen, P.; Naumann, F. Transforming Pairwise Duplicates to Entity Clusters for High-quality Duplicate Detection. *Journal of Data and Information Quality*, v. 12, n. 1, 2019.

Fu, J. et al. In-context Clustering-based Entity Resolution with Large Language Models: A Design Space Exploration. *Proceedings of the ACM on Management of Data*, v. 3, n. 4, 2025.

Huang, K. et al. ThriftLLM: On Cost-Effective Selection of Large Language Models for Classification Queries. *Proceedings of the VLDB Endowment*, v. 18, n. 11, p. 4410–4423, 2025.

Li, H. et al. On leveraging Large Language Models for Enhancing Entity Resolution. *arXiv preprint,* 2024.

Li, Y. et al. Deep Entity Matching with Pre-Trained Language Models. *Proceedings of the VLDB Endowment*, v. 14, n. 1, p. 50–60, 2020.

Nikoletos, K.; Ioannou, E.; Papadakis, G. The Five Generations of Entity Resolution on Web Data. In: STEFANIDIS, K. et al. (ed.). *Web Engineering*. Cham: Springer, 2024. p. 469–473.

Okayama, K.; Ito, H.; Morishima, A. Learning from Unknown-Unknowns: Inconsistency-Driven Sampling for Improving LLM Entity Matching. *Information Research*, 2026.

Pardo, F. de M. et al. GraLMatch: Matching Groups of Entities with Graphs and Language Models. In: SIMITSIS, A. et al. (ed.). *Proceedings of the 28th International Conference on Extending Database Technology (EDBT 2025).* Barcelona: OpenProceedings.org, 2025. p. 1–12.

Pardo, F. de M. et al. TransClean: Finding False Positives In Multi-Source Entity Matching Under Real-World Conditions Via Transitive Consistency. *IEEE Access*, v. 13, p. 195856–195870, 2025.

Peeters, R.; Bizer, C. Dual-Objective Fine-Tuning of BERT for Entity Matching. *Proceedings of the VLDB Endowment*, v. 14, n. 10, p. 1913–1921, 2021.

Peeters, R.; Bizer, C. Using ChatGPT for Entity Matching. In: **EUROPEAN CONFERENCE ON ADVANCES IN DATABASES AND INFORMATION SYSTEMS**. [S.l.]: Springer, 2023. p. 221–230.

Saeedi, A.; Peukert, E.; Rahm, E. Comparative Evaluation of Distributed Clustering Schemes for Multi-Source Entity Resolution. In*:* KIRIKOVA, M.; NØRVÅG, K.; PAPADOPOULOS, G. A. (ed.). *Advances in Databases and Information Systems*. Cham: Springer, 2017. p. 278–293.

Wang, T. et al. Match, Compare, or Select? An Investigation of Large Language Models for Entity Matching. In: *Proceedings of the 31st International Conference on Computational Linguistics*. 2025. p. 96–109.

ZEAKIS, A. et al. AvengER: Ensembling and Fine-Tuning LLMs for SELECT prompts in Entity Resolution. In: *Extended Semantic Web Conference*. 2025.