CANONWEAVE: PREVENTING COMMITMENT INTERFERENCE IN PARTIAL-ONTOLOGY SCHEMA RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

`city` and `product_name` can both resemble generic string fields to a local scorer, yet linking them forces incompatible ontology types and changes the feasible matches and blocking keys of their component. CanonWeave treats this *commitment interference* as a state-conditioned scheduling problem: the early and late policies jointly change prompt type context, queried edges, bundle volume, and solver initialization, while the method itself coordinates types and links before impact-aware LLM arbitration. Across six public data lakes, 27.7% of pre-resolution commitments enter a constraint conflict, and conflict/propagation increase from 15.2% / 1.4 in sparse components to 45.7% / 4.2 in dense ones, defining a verifier-placement phase map over coupling strength. The complete same-graph policy improves Match F1 by 3.5 points and ER F1 by 2.6 points while requesting 211 fewer bundles per lake; against prompt-first LLM-Chain, CanonWeave gains 5.1–9.7 Match-F1 and 3.2–5.2 ER-F1 points as mean FD/unit/temporal violations fall from 5.1/2.4/3.5% to 1.7/1.0/1.5%. Fixing queried edges and verifier outputs retains a 2.0-point gain from the coordinated initializer, isolating path dependence under identical evidence; a second resolver and Qwen2.5-72B preserve the state-conditioned advantage. The resulting design rule for partial-ontology lakes is direct: coordinate the canonical state before expensive local evidence acquires authority.

1 INTRODUCTION

The `city`/`product_name` collision in our evaluation illustrates a lake-wide decision hidden inside a plausible local match. Both columns contain capitalized strings, but the partial ontology assigns incompatible CITY and PRODUCT types. Accepting the edge would also constrain transitive closure, integrity penalties, and the blocking keys used by entity resolution (ER). A verifier acting on the unresolved pair therefore judges a different object from one acting on a feasible canonical assignment. Across six public lakes, conflict/propagation increase from 15.2% / 1.4 in components below 10 candidate edges to 45.7% / 4.2 above 40. We call this coupling *commitment interference*.

This failure matters because joins, analytics, governance, and record linkage inherit one shared type-and-link assignment Rahm & Do (2000); Doan et al. (2012); Bernstein et al. (2011); Christen (2012); Castro Fernandez et al. (2018); Hellerstein et al. (2017). Partial ontologies provide compatibility relations and integrity rules, but heterogeneous headers and values still leave residual ambiguity. Neural matchers and LLM pipelines strengthen pairwise evidence Konda et al. (2016); Li et al. (2020); Tu et al. (2023); Narayan et al. (2022); Parciak et al. (2024); Wang et al. (2026); Gungor et al. (2025); Liu et al. (2025); classical mapping and repair systems coordinate structure Miller et al. (2001); Fan & Geerts (2012); Chu et al. (2016); Rekatsinas et al. (2017). Neither line makes the intermediate canonical assignment the state that controls which learned judgments are requested and how they enter inference.

CanonWeave does. Candidate retrieval first constructs one graph over semantic types and schema links. A component-decomposed resolver then combines calibrated evidence with hard ontology compatibility and soft functional-dependency (FD), unit, and temporal penalties. Only after this joint solve does an impact–uncertainty gate select residual schema edges for LLM arbitration; verifier evidence is re-injected into the same objective and the affected component is solved from its

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

current feasible assignment. Learned evidence consequently refines a coordinated schema instead of becoming an independently binding local decision Garcez et al. (2019); Manhaeve et al. (2018).

The verifier-placement map in Section 6.2 separates three effects that are often conflated. The complete state-conditioned policy—resolved type context, residual selection, bundle volume, and coordinated initialization—improves Match F1 by 3.5 points and ER F1 by 2.6 points while using 211 fewer bundles per lake. With edge identities, bundle count, confidence values, and JSON outputs fixed byte-for-byte, coordinated initialization still contributes 2.0 Match-F1 points, directly measuring heuristic path dependence under fixed evidence. A priority-greedy resolver and Qwen2.5-72B preserve the qualitative result; candidate-exposure and symbolic-retention sweeps locate the regime in which coupling supplies leverage.

These mechanism controls accompany a strong end-to-end result. Against the fully specified prompt-first LLM-Chain, CanonWeave gains 5.1–9.7 Match-F1 and 3.2–5.2 ER-F1 points on every lake, while mean FD/unit/temporal violations fall from 5.1/2.4/3.5% to 1.7/1.0/1.5%. At identical per-lake bundle budgets, it also has the higher Match-F1 estimate in all 12/12 controlled LLMATCH- and Schemora-adapter comparisons. The technical contribution is therefore both an inference design—joint canonicalization followed by impact-aware residual arbitration—and a diagnostic account that connects commitment interference to canonical quality, consistency, downstream ER, and verifier demand.

2 Related Work

Classical schema matching, mapping, and repair coordinate heterogeneous sources through correspondences, dependencies, and global consistency Rahm & Do (2000); Doan et al. (2012); Bernstein et al. (2011); Miller et al. (2001); Fan & Geerts (2012); Chu et al. (2016); Rekatsinas et al. (2017). Valentine standardizes matcher evaluation, while Aurum, Seeping Semantics, and Starmie retrieve related columns or tables Koutras et al. (2021); Castro Fernandez et al. (2018); Fernandez et al. (2018); Fan et al. (2023). CanonWeave builds on these structural ideas but treats the *intermediate joint assignment* as an inference state: it determines both which residuals merit learned evidence and the feasible component into which that evidence is inserted.

Neural matchers and table models improve representations, pair scores, prompts, and evidence aggregation Li et al. (2020); Wang et al. (2023); Tu et al. (2023); Narayan et al. (2022); Li et al. (2024); Parciak et al. (2024); Bopardikar et al. (2025); Peeters et al. (2025). Recent LLM systems make local reasoning economical in different ways: Magneto couples small-model retrieval with LLM reranking, LLMATCH composes table selection and column alignment, and Schemora combines metadata enrichment, hybrid retrieval, selection, and ranking Liu et al. (2025); Wang et al. (2026); Gungor et al. (2025). Their retrieval and model-use units define distinct end-to-end operating points. Our LLMATCH- and Schemora-class comparisons instead isolate the canonicalization decision under common candidate support, verifier family, bundle count, and downstream evaluation; all claims about those rows are adapter-level.

Active entity matching allocates labels to informative pairs, and selective prediction trades coverage for risk Meduri et al. (2020); Geifman & El-Yaniv (2017). CanonWeave adds global consequence to uncertainty: force-flipping an edge measures how much it can change the coordinated objective, and accepted evidence re-enters the constrained resolver. This inference-time coupling is related to decision-focused and neural–symbolic learning Elmachtoub & Grigas (2022); Garcez et al. (2019); Manhaeve et al. (2018), but the paper's object is placement within a deterministic canonicalization path. Commitment-interference traces, identical-output injection, an independent priority-greedy resolver, and a second verifier distinguish the complete scheduling policy from its initialization-conditioned component.

3 Problem Setup

Our data lake comprises heterogeneous tables $\mathcal{T} = \{T_1, \ldots, T_n\}$. For every column c, the available information consists of its header $h(c)$, value samples $V(c)$, and candidate semantic types taken from a partial ontology $\mathcal{O}$. This ontology specifies type compatibility through Γ and supplies integrity constraints $\mathcal{K}$, including functional dependencies, rules for units, and invariants over time. These

constraints are incomplete but useful: CanonWeave requires type compatibility for admissibility, while applying soft penalties to the other integrity constraints to permit integration of noisy lakes.

Semantic typing and schema matching are inferred jointly by the resolver, with entity resolution following downstream. Specifically, each column receives a type $s(c) \in \mathcal{S} \cup \{\perp\}$, and columns representing the same attribute receive predicted schema links $(c_i, c_j) \in \mathcal{M}$; the resulting resolved schema then determines the canonical attributes and the blocking keys used in record-level ER. A typed attribute graph encodes the candidate decisions about types and matches, with edge evidence scored using headers and values alongside embeddings and simple statistics.

The resolved state is the joint assignment $Z = (x, y)$ returned by the resolver. Its hard invariant is membership in the feasible set $\mathcal{F}(\Gamma)$: one type per column, symmetric and transitive links, and ontology-compatible linked types. FD, unit, and temporal checks shape the selected feasible state through penalties and are reported as empirical consistency outcomes. A verifier judgment can change scores or fix a compatible residual decision, but the resulting assignment becomes canonical only after the same resolver maps it back into $\mathcal{F}(\Gamma)$. This enforcement boundary makes evidence-injection order a well-defined intervention.

For the mechanism analysis, a *pre-resolution commitment* is a verifier decision inserted before the first joint solve. We mark it as conflicting when the first global pass associates that edge with either a violated hard condition or a positive FD, unit, or temporal slack; an edge is one denominator even if it activates several checks. An *overturn* occurs when the repaired assignment changes that committed edge itself. Its propagation count is the number of other type or match decisions in the affected component that differ between the committed and repaired assignments. We average conflict and overturn over distinct commitments and propagation over overturned commitments, preserving component dependence in the bootstrap.

Let E_{type} be candidate typing edges and E_{match} candidate schema-match edges. Stage II solves

$$(x^*, y^*) \;=\; \arg\max_{x,y} \sum_{(c,s)\in E_{\text{type}}} w^{\text{type}}_{c,s} x_{c,s} + \sum_{(i,j)\in E_{\text{match}}} w^{\text{match}}_{i,j} y_{i,j} - \sum_{q\in\mathcal{Q}} \beta_q \, p_q,$$

with constraints requiring a single type for each column, compatibility of schema links with the ontology, and penalty slacks p_q that are non-negative and derived from sampled checks of FDs, units, and temporal properties. The solution defines the canonical schema consumed by downstream ER, isolating the lake-level state through which typing and linking errors propagate into record decisions. CanonWeave computes this state with the linearized relaxation, feasibility-preserving rounding, and local refinement described in Section 4.

4 CanonWeave: Joint Resolution and Residual Arbitration

CanonWeave treats semantic types, schema links, and verifier evidence as inputs to one canonicalization program. Candidate scores propose decisions, ontology compatibility limits their joint support, integrity checks penalize incoherent combinations, and LLM judgments update only components whose solutions remain uncertain and consequential.

Let $R(G, A; Z_{\text{init}}, \Gamma, \mathcal{K})$ denote the deterministic resolver for graph G, verifier evidence A, compatibility relation Γ, integrity checks $\mathcal{K}$, and initializer Z_{init}. CanonWeave computes $Z_0 = R(G, \emptyset; \emptyset, \Gamma, \mathcal{K})$, queries A_H on residual set $H(Z_0)$, and returns $Z_1 = R(G, A_H; Z_0, \Gamma, \mathcal{K})$ after affected-component re-optimization. The early schedule derives local-argmax type context and A_{pre} from the unresolved state, then computes $R(G, A_{\text{pre}}; \emptyset, \Gamma, \mathcal{K})$. Thus the complete policy intervention changes four linked quantities: prompt type context, selected residuals, bundle count, and initializer. The same-output intervention sets $A_{\text{pre}} = A_H$, including edge identities, JSON decisions, confidence, and count; it isolates initializer-conditioned optimization-path sensitivity for the tested resolver.

Figure 1 shows the key authority boundary: verifier evidence can change a canonical state, but does not replace its constraints or decision variables.

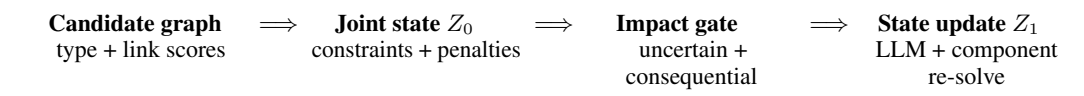


Figure 1: CanonWeave makes the coordinated assignment Z_0, rather than an isolated pair score, the context for residual selection and LLM arbitration.

4.1 Stage I: Candidate Generation

For each column c, Stage I forms a 1,152-dimensional neural representation by concatenating a 384-dimensional MiniLM header embedding with a 768-dimensional `roberta-base` value embedding over 50 stratified cells Wang et al. (2020). Value sketches—MinHash for strings, quantiles for numerics, and periodicity for timestamps—enter as additional scorer features. HNSW retrieval proposes top-$k_s = 15$ typing and top-$k_m = 25$ schema-match candidates after coarse filtering Fan et al. (2023). Appendix Table 11 quantifies gold-edge exposure and the normalized level at which coordination loses sufficient support.

4.2 Stage II: Constraint-Aware Global Resolver

Stage II makes a joint choice of semantic types and schema links, enforcing ontology compatibility and applying soft penalties for integrity. The linearized objective used in our relax-and-round solve is

$$\max_{x,y,z,p} \left(\sum_{c,s} w_{c,s}^{\text{type}} x_{c,s} + \sum_{i,j} w_{i,j}^{\text{match}} y_{i,j} - \sum_{q \in \mathcal{Q}} \beta_q p_q \right), \tag{1}$$

where a type assignment is encoded by $x_{c,s}$, a schema match by $y_{i,j}$, and compatibility auxiliaries by z; normalized slacks $p_q \geq 0$ measure violations in sampled checks of FDs, units, and temporal properties. The weights are calibrated log-odds from the fixed Stage I type and match scorers, so evidence terms and penalties share a stable scale across lakes. Hard constraints enforce one type per column, symmetric and transitive schema links, and ontology-compatible matches; soft penalties allow noisy lakes to violate imperfect rules without collapsing the entire solution. Appendices A.1 and A.2 give the complete linearization, score construction, and deterministic local-search rule.

The main resolver solves the continuous relaxation with ECOS through CVXPY, greedily rounds x and y under the hard constraints, and runs deterministic 2-opt refinement within each connected component. The initial pass starts from the fixed candidate order; a residual update initializes the same rounding and local-improvement rules from the current feasible component assignment. We fix $\beta_{\text{FD}} = 2.0$, $\beta_{\text{unit}} = 3.0$, and $\beta_{\text{temp}} = 1.5$ across datasets. On 50 exact instances solved by Gurobi within 10 minutes, this procedure reaches an objective within 2.3% of the ILP optimum on average.

The independent priority-greedy resolver in Appendix Table X2 avoids continuous relaxation. It sorts candidate type and match decisions by evidence score minus the incremental soft-penalty cost, admits the highest-scoring feasible decision, breaks ties by larger constraint impact and then stable edge identifier, and repairs each connected component by removing its lowest-margin conflicting decision before reconsidering skipped candidates. This different optimization path preserves the same hard compatibility and penalty semantics while testing whether the ordering result depends on relax-and-round dynamics.

4.3 Stage III: Selective LLM Verification as Residual Arbitration

LLM verification acts as residual arbitration rather than front-end evidence gathering. Its use is restricted to the point after resolver enforcement of the strongest symbolic signal available. Here an edge always means $e \in E_{\text{match}}$. If its calibrated Stage I match probability is p_e, the binary certainty margin is $\Delta_e = |p_e - (1 - p_e)|$. To compute impact, we force e to the opposite value, run the same deterministic feasibility repair and component re-optimization, and measure the absolute change $I_e = |J(Z_0) - J(Z_0^{\neg e})|$ in Equation 1. Edges with high impact ($I_e > 0.05$) and low certainty ($\Delta_e < 0.15$) are sent to the verifier, which selects 2.1–4.9% of candidate match edges over the operating profile reported for each dataset. We use GPT-4-Turbo (`gpt-4-1106-preview`) as the verifier, supplying structured prompts with headers, 10 sampled values, and context from the

ontology; early prompts use Stage I local-argmax types, whereas late prompts use the types in Z_0. Its JSON output contains the decision plus confidence, a rationale, and quoted evidence. Self-consistency uses 3-way evaluation at $T \in \{0.0, 0.2, 0.4\}$.

One verification bundle is one logical edge decision. A cache miss issues the three API invocations above; a cache hit issues none. Bundle counts include all logical requests, counted before cache lookup; token quantiles are computed per executed API invocation, and P95 latency is wall-clock minutes per complete lake run across seeds. A compatible unanimous decision is fixed provisionally for the affected component solve, while a two-of-three majority becomes a unit-scale signed log-odds term; the reported JSON confidence is retained for provenance and exact replay but does not rescale this term. An ontology-incompatible unanimous decision is also softened. During deterministic repair, a provisional fix is released exactly when it makes the component infeasible or when the feasible component solution obtained after release has a strictly higher full objective after integrity penalties; ties retain the fix. Such a release is the overturn measured in Section 3. If no binary majority emerges, CanonWeave keeps the resolver decision.

The main operating point uses this verifier, prompt template, self-consistency policy, and bundle definition; the all-six Qwen2.5-72B experiment tests whether the ordering persists under a second verifier Yang et al. (2024).

4.4 Downstream Entity Resolution

Every method feeds the same downstream ER computation: identical blocking logic, record matcher, confidence threshold, and low-confidence policy. Resolved types and constraints define blocking keys, cutting comparisons by 94–99% relative to all-pairs search; remaining pairs are scored by a two-layer MLP over concatenated `roberta-base` attribute embeddings Papadakis et al. (2020); Li et al. (2020). ER F1 therefore measures the downstream consequence of the upstream canonical state under controlled record-level computation.

5 Experimental Setup

The evaluation asks three questions: whether commitment interference is measurable (RQ1), whether state-conditioned arbitration improves canonical and downstream outcomes under controlled computation (RQ2), and how candidate support, symbolic structure, optimizer, and verifier family move that advantage (RQ3).

5.1 Benchmark Construction and Split Isolation

The public core contains six evaluation assemblies spanning 24–3,215 tables, 486–18,721 columns, 28–62 ontology types, and 18–89 FD rules per lake. Table 1 specifies how each source becomes a canonicalization task; exact candidate, reference-link, and consistency-check denominators appear in Appendix Tables 8 and 11.

Released labels are used where available; declared schema and key correspondences define references for joined or generated assemblies. Before model fitting, source-key uniqueness, reciprocal correspondence, and type compatibility validate those references. Ontology nodes and compatibility edges are frozen from public schema descriptions and canonical field maps; FD rules come from declared keys and dependencies, while unit and temporal checks come from typed source metadata. No test outcome selects these structures.

Complete source tables, their columns, and incident record pairs are assigned together. We use official partitions when supplied and otherwise stratify tables 70/10/20 by source and schema family under each seed. Candidate scorers fit on training data, Platt calibration and thresholds use validation, and components are built only after assignment so none crosses folds. Test labels enter neither retrieval, ontology construction, gating, nor penalty selection.

5.2 Baselines and Controlled Comparators

LLM-Chain is the uniform prompt-first anchor. On the fixed Stage I graph, it forms the unresolved initialization from local-argmax type and match decisions, then applies the same binary certainty

Table 1: Public-core task construction. Source tables are the split unit under the common table-disjoint protocol.

LAKE	CANONICALIZATION TASK	REFERENCE DECISIONS
WDC Products	Group offers by source table; align product fields Primpeli et al. (2019)	Released product IDs and canonical fields
WebTables	Align relational-table columns to benchmark canonical columns Cafarella et al. (2008)	Released column correspondences
Flights+ Weather	Canonicalize BTS flight and NOAA station-hour fields before key-based joining U.S. Department of Transportation, Bureau of Transportation Statistics (2026); NOAA National Centers for Environmental Information (2026)	Declared fields, source keys, joined identities
OSM Slices	Align tags across nonoverlapping Geofabrik geographic tables OpenStreetMap contributors & Geofabrik GmbH (2026)	OSM tags, feature IDs, canonical field map
MAG Profiles	Align author, paper, venue, and affiliation profile fields Sinha et al. (2015)	Released IDs and canonical profile fields
Retail Lakehouse	Align heterogeneous source tables reshaped from TPC-DS relations Transaction Processing Performance Council (2024)	Declared schema, keys, generated identities

margin and force-flip impact operator defined in Section 4.3 to that state before any joint solve. Candidate edges are visited by decreasing impact and then stable edge identifier; every edge satisfying $I > 0.05$ and $\Delta < 0.15$ receives the same three-invocation prompt bundle as Stage III. Unanimous compatible outputs become provisional fixes, majority outputs become unit-scale log-odds terms, and ontology-incompatible unanimous outputs are softened exactly as in CanonWeave. After all gate-qualified edges are processed, LLM-Chain executes one Stage II resolve-and-repair pass and requests no residual bundles. This defines its query order, stopping rule, and output-to-repair mapping.

Table 2 includes LLM-Chain for every dataset and two prespecified additional references: WDC/MAG use NeuralMatch and Ditto; Web/Retail use NeuralMatch and LLM-Prompt; Flights uses Rule+Sim and NeuralMatch; OSM uses Rules+Global and LLM-Prompt. NeuralMatch represents transformer matching Konda et al. (2016); Mudgal et al. (2018); Brunner & Stockinger (2020); Li et al. (2020); Peeters & Bizer (2021); Ditto is the system described by Li et al. (2020). LLM-Prompt applies the common verifier independently to candidate edges without chain context. Rule+Sim ranks candidate links by rule and similarity scores without joint resolution, whereas Rules+Global passes those scores through the Stage II constraints while excluding neural and LLM evidence. For these link-only baselines, the fixed Stage I local-argmax types enter the common canonical repair and downstream ER; only the schema-link predictor varies. The "-LLM" ablation retains all Stage I learned scores and removes only residual verification. LLM-Everywhere verifies every candidate edge as an exhaustive-use reference.

The controlled frontier adapts LLMATCH and Schemora modules to the shared candidate graph, verifier family, downstream ER, and exact per-dataset bundle budget Wang et al. (2026); Gungor et al. (2025). It compares adapter-level outcomes while retaining architecture, ranking, and prompt differences (Appendix A.2.11). Magneto's native small-model retrieval and LLM reranking jointly change candidate support and the model-use unit Liu et al. (2025); it therefore informs the neighboring full-system design space rather than this matched placement estimand.

5.3 METRICS AND PROTOCOL

We report macro-Type F1 including $\perp$, pairwise Match F1, downstream pairwise ER F1, post-integration FD/unit/temporal violation rates, verified-edge fraction, verification-bundle count, per-invocation token quantiles, and P95 lake-run latency. Verified-edge fraction is the number of unique logically queried edges divided by $|E_{\text{match}}|$. One bundle is one logical edge decision: it comprises three temperature-specific invocations on a cache miss and no new invocation on a cache hit, while bundle counts include both because they are recorded before lookup. Bundle count is the primary model-use measure because it is invariant to provider pricing; every budget-matched comparison fixes this count exactly. Candidate generation uses `all-MiniLM-L6-v2`, `roberta-base`, and HNSW; FD checks and downstream blocking are shared across methods Wang et al. (2020); Papenbrock et al. (2015); Papadakis et al. (2020).

Main comparisons use five fixed seeds (42, 73, 101, 137, and 211); scalar scores average seeds, and public-core aggregates are unweighted means across the six datasets. We construct 95% intervals from 10K paired hierarchical bootstrap replicates. Each replicate resamples seed runs and then tables for Type F1, connected candidate components for Match F1 and violations, or entity blocks for ER F1. Detailed tables include candidate, gold-edge, and consistency-check denominators. Runs use Ubuntu 22.04, Python 3.11.7, PyTorch 2.2.2, Transformers 4.45.0, two Intel Xeon Gold 6248R CPUs, 512GB RAM, and one NVIDIA A100 80GB GPU.

5.4 CONTROLLED INTERVENTIONS

For RQ1–RQ2, the paired-policy intervention fixes inputs, graph, resolver, verifier, token cap, gates, seeds, and ER stack, then changes the four schedule-linked quantities identified in Section 4. Queried edges and bundle count are policy outcomes. Identical-output injection additionally fixes edges, cached JSON, confidence, and count, leaving initializer-conditioned solver path as the treatment. For RQ3, priority-greedy changes the optimizer, Qwen2.5-72B changes the verifier, component ablations remove ontology compatibility, soft penalties, or residual verification, and candidate-exposure and symbolic-retention sweeps vary the structure available to both schedules.

6 RESULTS

6.1 RQ1: COMMITMENT INTERFERENCE

The first global pass exposes the mechanism: 27.7% of pre-resolution verifier commitments participate in a compatibility or integrity conflict, repair overturns 64.6% of those edges, and each overturn changes 2.6 additional decisions on average. Conflict/propagation rise from 15.2% / 1.4 below 10 candidate edges to 45.7% / 4.2 above 40 (Appendix Table 9). This density gradient connects the `price_usd`/`price_eur` collision to a general component-level phenomenon.

The lake-level consequence is consistent: Table 2 shows 5.1–9.7 Match-F1 and 3.2–5.2 ER-F1 gains over LLM-Chain across all six datasets. Over 52,850 checks, Table 3 reduces mean FD/unit/temporal violations by 3.4/1.4/2.0 points; all paired intervals exclude zero.

Component ablations reinforce the account (Appendix Table 14): removing ontology compatibility, soft penalties, or residual arbitration lowers Match F1 by 4.8–7.3, 2.5–6.3, or 3.7–4.6 points, respectively. Global structure supplies the feasible substrate; sparse learned evidence resolves what remains.

6.2 RQ2: VERIFIER-PLACEMENT MAP

The complete policy comparison fixes inputs, graph, splits, seeds, GPT-4-Turbo, token cap, gates, resolver, ER stack, and bootstrap units. It changes unresolved versus resolved type context, state-conditioned edge selection, resulting bundle volume, and initializer. CanonWeave has higher quality on all six lakes: +3.5 Match-F1 and +2.6 ER-F1 points on average, with 211 fewer bundles. Appendix Tables 8 and 9 report per-lake outcomes and paired intervals.

Table 2: Public-core quality under the shared downstream ER protocol. Each block includes LLM-Chain and two prespecified references from Section 5.2.

DATASET	METHOD	TYPE F1	MATCH F1	ER F1
WDC Products	NeuralMatch	0.742	0.684	0.847
	Ditto	0.768	0.712	0.862
	LLM-Chain	0.793	0.738	0.881
	CanonWeave	0.851	0.794	0.914
WebTables	NeuralMatch	0.698	0.621	0.793
	LLM-Prompt	0.754	0.673	0.824
	LLM-Chain	0.772	0.691	0.841
	CanonWeave	0.823	0.752	0.878
Flights+Weather	Rule+Sim	0.812	0.698	0.834
	NeuralMatch	0.834	0.724	0.856
	LLM-Chain	0.867	0.768	0.884
	CanonWeave	0.904	0.831	0.921
OSM Slices	Rules+Global	0.798	0.721	0.854
	LLM-Prompt	0.821	0.742	0.871
	LLM-Chain	0.838	0.761	0.886
	CanonWeave	0.876	0.812	0.918
MAG Profiles	NeuralMatch	0.723	0.642	0.821
	Ditto	0.751	0.674	0.843
	LLM-Chain	0.778	0.702	0.864
	CanonWeave	0.842	0.768	0.902
Retail Lakehouse	NeuralMatch	0.684	0.591	0.778
	LLM-Prompt	0.734	0.652	0.812
	LLM-Chain	0.768	0.684	0.841
	CanonWeave	0.847	0.781	0.893

Table 3: Unweighted means across the six public-core datasets. Parentheses give 95% paired hierarchical-bootstrap intervals for the reduction achieved by CanonWeave.

METHOD	FD ↓	UNIT ↓	TEMPORAL ↓
LLM-Chain	5.1%	2.4%	3.5%
CanonWeave	1.7%	1.0%	1.5%
Reduction	3.4 (3.0, 3.8)	1.4 (1.1, 1.7)	2.0 (1.6, 2.4)

Retail combines the highest commitment-conflict rate (42.0%) with the largest policy gain (5.4 Match-F1 points); OSM's smaller components produce a near-tie under the independent resolver. The effect therefore tracks interaction density rather than a uniform threshold shift.

Identical-output injection removes selection and verifier volume as explanations. Starting the same relax-and-round heuristic from the coordinated assignment improves all 6/6 datasets by 2.0 Match-F1 and 1.5 ER-F1 points on average, with a 0.65-point mean-violation reduction; all paired intervals exclude zero (Appendix Table X1). This is direct evidence of initialization-conditioned path dependence, while agreement with the complete policy and priority-greedy comparisons supports the implemented state-conditioned design across two optimization procedures.

6.3 RQ3: OPTIMIZER, VERIFIER, AND SUPPORT REGIME

Priority-greedy changes scoring, tie-breaking, and repair dynamics while preserving ontology and integrity semantics. State-conditioned arbitration improves 5/6 datasets by 2.2 Match-F1 and 1.7 ER-F1 points on average, with 73 fewer bundles (Appendix Table X2).

At CanonWeave's exact per-dataset bundle budgets, it has the higher Match-F1 estimate in all 12/12 controlled LLMATCH/Schemora adapter cells (Appendix Table 10). Mean Match-F1 advantages

Table 4: Verifier-placement map. Effects are unweighted public-core means; positive quality differences favor state-conditioned residual arbitration.

CONTROL	EVIDENCE AND STATE HELD FIXED	MATCH / ER EFFECT	CONSISTENCY / EFFICIENCY
Identical output	Edges, bundles, JSON, confidence; initializer changes	+2.0 / +1.5; 6/6	-0.65 mean violation; bundles equal
Complete policy	Graph, resolver, verifier, gates, ER; state policy changes	+3.5 / +2.6; 6/6	-0.9/-0.5/-0.7 violations; -211 bundles
Priority-greedy	Graph and verifier; optimizer family changes	+2.2 / +1.7; 5/6	-0.5/-0.3/-0.3 violations; -73 bundles
Qwen2.5-72B	Graph and resolver; verifier family changes	+2.9 / +2.3; 6/6	-0.8/-0.4/-0.6 violations; -181 bundles

are +2.1 / +1.4 points, with paired macro intervals [+1.5,+2.8] / [+0.7,+2.0]. These results establish the adapter-level ranking for canonicalization under common candidate support, verifier family, and bundle count.

With Qwen2.5-72B-Instruct substituted under the same graph, prompt schema, token cap, gates, splits, and ER stack, the advantage holds on 6/6 datasets: 2.9 Match-F1 and 2.3 ER-F1 points with 181 fewer bundles per lake (Appendix Table X3).

Support regime. Candidate loss is gradual before the fitted public-core bend around 0.84 of baseline recovered exposure (Appendix Table 11). Removing ontology edges and integrity rules contracts the margin: interpolation places its half-margin near 0.72 retained structure, and the mean direction remains positive at the lowest evaluated 0.60 retained structure (Appendix Table 13). These descriptive coordinates match the density evidence: sparse OSM is the second-resolver near-tie, whereas dense MAG and Retail show stronger conflict and propagation.

The evaluated regime comprises partial-ontology lakes with high candidate exposure and enough compatibility and integrity structure to couple ambiguous edges. As that structure becomes sparse, fewer interactions remain and the placement margin contracts toward a tie. Across the six public lakes and tested perturbations, residual arbitration retains its direction across two resolvers and two verifier families; within this regime, it improves canonical quality, downstream ER, consistency, and bundle efficiency together.

7 CONCLUSION

CanonWeave makes the coordinated canonical state the authority boundary for learned evidence. This design addresses a measurable failure of local-first verification: 27.7% of pre-resolution commitments enter a constraint conflict, and each overturned conflict changes 2.6 neighboring decisions on average. Resolving types and links before impact-aware arbitration converts those interactions into simultaneous gains in schema matching, downstream ER, FD/unit/temporal consistency, and verification efficiency across six public lakes.

The evidence distinguishes the complete state-conditioned policy from its solver-path component. The same-graph intervention yields +3.5 Match-F1 and +2.6 ER-F1 points with fewer bundles, while identical queried edges and outputs retain a 2.0-point Match-F1 gain from coordinated initialization. Priority-greedy and Qwen2.5-72B preserve the qualitative result, and exact-budget adapters plus structure-degradation sweeps locate its competitive and operating regime. The resulting design rule is direct: in partial-ontology canonicalization, first coordinate the state that links must jointly satisfy, then spend verifier capacity on the residual decisions that can change it.

Ethics Statement

Canonical integration can improve consistency and governance, but an incorrect match can also propagate privacy or access-control errors across joined data. The central evaluation uses six public benchmarks. In operational use, verifier prompts should exclude direct identifiers, canonical changes should retain provenance, and downstream joins should support review and rollback.

Reproducibility Statement

Dataset preparation, shared downstream evaluation, resolver and verifier protocols, hyperparameters, uncertainty procedure, compute environment, and sensitivity analyses are specified in Sections 3–6 and the appendix.

AI Use Statement

Generative AI tools supported revisions to the manuscript's wording, organization, title, and reference formatting during manuscript preparation.

References

Philip A Bernstein, Jayant Madhavan, and Erhard Rahm. Generic schema matching, ten years later. *Proceedings of the VLDB Endowment*, 4(11):695–701, 2011. doi: 10.14778/3402707.3402710.

Rohan Bopardikar, Jin Wang, and Jia Zou. Structured multi-step reasoning for entity matching using large language model, 2025. URL https://arxiv.org/abs/2511.22832.

Ursin Brunner and Kurt Stockinger. Entity matching with transformer architectures—a step forward in data integration. In *Proceedings of EDBT*, pp. 463–473, 2020. doi: 10.5441/002/edbt.2020.58.

Michael J Cafarella, Alon Halevy, Daisy Zhe Wang, Eugene Wu, and Yang Zhang. Webtables: Exploring the power of tables on the web. *Proceedings of the VLDB Endowment*, 1(1):538–549, 2008. doi: 10.14778/1453856.1453916.

Raul Castro Fernandez, Ziawasch Abedjan, Famien Koko, Gina Yuan, Samuel Madden, and Michael Stonebraker. Aurum: A data discovery system. In *2018 IEEE 34th International Conference on Data Engineering*, pp. 1001–1012, 2018. doi: 10.1109/ICDE.2018.00094.

Peter Christen. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Data-Centric Systems and Applications. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-31164-2. doi: 10.1007/978-3-642-31164-2.

Xu Chu, Ihab F Ilyas, Sanjay Krishnan, and Jiannan Wang. Data cleaning: Overview and emerging challenges. In *Proceedings of the 2016 International Conference on Management of Data*, pp. 2201–2206. ACM, 2016. doi: 10.1145/2882903.2912574.

AnHai Doan, Alon Halevy, and Zachary G. Ives. *Principles of Data Integration*. Morgan Kaufmann, 2012. ISBN 978-0-12-416044-6. doi: 10.1016/C2011-0-06130-6.

Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68(1): 9–26, 2022. doi: 10.1287/mnsc.2020.3922.

Grace Fan, Jin Wang, Yuliang Li, Dan Zhang, and Renée J. Miller. Semantics-aware dataset discovery from data lakes with contextualized column-based representation learning. *Proceedings of the VLDB Endowment*, 16(7):1726–1739, 2023. doi: 10.14778/3587136.3587146.

Wenfei Fan and Floris Geerts. *Foundations of Data Quality Management*. Synthesis Lectures on Data Management. Morgan & Claypool, 2012. doi: 10.1007/978-3-031-01892-3.

Raul Castro Fernandez, Essam Mansour, Abdulhakim A. Qahtan, Ahmed K. Elmagarmid, Ihab F. Ilyas, Samuel Madden, Mourad Ouzzani, Michael Stonebraker, and Nan Tang. Seeping semantics: Linking datasets using word embeddings for data discovery. In *Proceedings of IEEE ICDE*, pp. 989–1000, 2018. doi: 10.1109/ICDE.2018.00093.

Artur d'Avila Garcez, Marco Gori, Luis C. Lamb, Luciano Serafini, Michael Spranger, and Son N. Tran. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. *Journal of Applied Logics*, 6(4):611–631, 2019. URL https://openaccess.city.ac.uk/id/eprint/22984/.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems 30*, pp. 4878–4887, 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Osman Erman Gungor, Derak Paulsen, and William Kang. SCHEMORA: Schema matching via multi-stage recommendation and metadata enrichment using off-the-shelf LLMs, 2025. URL https://arxiv.org/abs/2507.14376.

Joseph M. Hellerstein, Vikram Sreekanti, Joseph E. Gonzalez, James Dalton, Akon Dey, Sreyashi Nag, Krishna Ramachandran, Sudhanshu Arora, Arka Bhattacharyya, Shirshanka Das, Mark Donsky, Gabe Fierro, Chang She, Carl Steinbach, Venkat Subramanian, and Eric Sun. Ground: A data context service. In *8th Biennial Conference on Innovative Data Systems Research*, 2017. URL https://www.cidrdb.org/cidr2017/papers/p111-hellerstein-cidr17.pdf.

Pradap Konda, Sanjib Das, Paul Suganthan G. C., AnHai Doan, Adel Ardalan, Jeffrey R. Ballard, Han Li, Fatemah Panahi, Haojun Zhang, Jeffrey F. Naughton, Shishir Prasad, Ganesh Krishnan, Rohit Deep, and Vijay Raghavendra. Magellan: Toward building entity matching management systems. *Proceedings of the VLDB Endowment*, 9(12):1197–1208, 2016. doi: 10.14778/2994509.2994535.

Christos Koutras, George Siachamis, Andra Ionescu, Kyriakos Psarakis, Jerry Brons, Marios Fragkoulis, Christoph Lofi, Angela Bonifati, and Asterios Katsifodimos. Valentine: Evaluating matching techniques for dataset discovery. In *Proceedings of IEEE ICDE*, pp. 468–479, 2021. doi: 10.1109/ICDE51399.2021.00047.

Peng Li, Yeye He, Dror Yashar, Weiwei Cui, Song Ge, Haidong Zhang, Danielle Rifinski Fainman, Dongmei Zhang, and Surajit Chaudhuri. Table-gpt: Table fine-tuned gpt for diverse table tasks. *Proceedings of the ACM on Management of Data*, 2(3):1–28, 2024. doi: 10.1145/3654979.

Yuliang Li, Jinfeng Li, Yoshihiko Suhara, AnHai Doan, and Wang-Chiew Tan. Deep entity matching with pre-trained language models. *Proceedings of the VLDB Endowment*, 14(1):50–60, 2020. doi: 10.14778/3421424.3421431.

Yurong Liu, Eduardo H. M. Peña, Aécio Santos, Eden Wu, and Juliana Freire. Magneto: Combining small and large language models for schema matching. *Proceedings of the VLDB Endowment*, 18 (8):2681–2694, 2025. doi: 10.14778/3742728.3742757.

Robin Manhaeve, Sebastijan Dumančić, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. Deepproblog: Neural probabilistic logic programming. In *Advances in Neural Information Processing Systems 31*, pp. 3753–3763, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/dc5d637ed5e62c36ecb73b654b05ba2a-Abstract.html.

Venkata Vamsikrishna Meduri, Lucian Popa, Prithviraj Sen, and Mohamed Sarwat. A comprehensive benchmark framework for active learning methods in entity matching. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pp. 1133–1147, 2020. doi: 10.1145/3318464.3380597.

Renée J. Miller, Mauricio A. Hernández, Laura M. Haas, Lingling Yan, C. T. Howard Ho, Ronald Fagin, and Lucian Popa. The clio project: Managing heterogeneity. *ACM SIGMOD Record*, 30(1): 78–83, 2001. doi: 10.1145/373626.373713.

Sidharth Mudgal, Han Li, Theodoros Rekatsinas, AnHai Doan, Youngchoon Park, Ganesh Krishnan, Rohit Deep, Esteban Arcaute, and Vijay Raghavendra. Deep learning for entity matching: A design space exploration. In *Proceedings of ACM SIGMOD*, pp. 19–34, 2018. doi: 10.1145/3183713.3196926.

Avanika Narayan, Ines Chami, Laurel Orr, and Christopher Ré. Can foundation models wrangle your data? *Proceedings of VLDB*, 16(4):738–746, 2022. doi: 10.14778/3574245.3574258.

NOAA National Centers for Environmental Information. Global hourly: Integrated surface database (ISD), 2026. URL https://www.ncei.noaa.gov/products/land-based-station/integrated-surface-database. Dynamic dataset; accessed 2026-08-30.

OpenStreetMap contributors and Geofabrik GmbH. Openstreetmap data extracts. Geofabrik Download Server, 2026. URL https://download.geofabrik.de/. Accessed 2026-08-30.

George Papadakis, Dimitrios Skoutas, Emmanouil Thanos, and Themis Palpanas. Blocking and filtering techniques for entity resolution: A survey. *ACM Computing Surveys*, 53(2):1–42, 2020. doi: 10.1145/3377455.

Thorsten Papenbrock, Jens Ehrlich, Jannik Marten, Tommy Neubert, Jan-Peer Rudolph, Martin Schönberg, Jakob Zwiener, and Felix Naumann. Functional dependency discovery: An experimental evaluation of seven algorithms. *Proceedings of the VLDB Endowment*, 8(10):1082–1093, 2015. doi: 10.14778/2794367.2794377.

Marcel Parciak, Brecht Vandevoort, Frank Neven, Liesbet M. Peeters, and Stijn Vansummeren. Schema matching with large language models: An experimental study. In *Proceedings of Workshops at the 50th International Conference on Very Large Data Bases*. VLDB Endowment, 2024. URL https://vldb.org/workshops/2024/proceedings/TaDA/TaDA.8.pdf.

Ralph Peeters and Christian Bizer. Dual-objective fine-tuning of bert for entity matching. *Proceedings of the VLDB Endowment*, 14(10):1913–1921, 2021. doi: 10.14778/3467861.3467878.

Ralph Peeters, Aaron Steiner, and Christian Bizer. Entity matching using large language models. In *Proceedings of the 28th International Conference on Extending Database Technology*, pp. 529–541. OpenProceedings.org, 2025. doi: 10.48786/edbt.2025.42.

Anna Primpeli, Ralph Peeters, and Christian Bizer. The wdc training dataset and gold standard for large-scale product matching. In *Companion Proceedings of the 2019 World Wide Web Conference*, pp. 381–386, 2019. doi: 10.1145/3308560.3316609.

Erhard Rahm and Hong Hai Do. Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin*, 23(4):3–13, 2000. URL https://sites.computer.org/debull/A00dec/A00DEC-CD.pdf.

Theodoros Rekatsinas, Xu Chu, Ihab F Ilyas, and Christopher Ré. Holoclean: Holistic data repairs with probabilistic inference. *Proceedings of the VLDB Endowment*, 10(11):1190–1201, 2017. doi: 10.14778/3137628.3137631.

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th International Conference on World Wide Web*, pp. 243–246. ACM, 2015. doi: 10.1145/2740908.2742839.

Transaction Processing Performance Council. *TPC Benchmark DS: Standard Specification, Version 4.0.0*. Transaction Processing Performance Council, November 2024. URL https://www.tpc.org/TPC_Documents_Current_Versions/pdf/TPC-DS_v4.0.0.pdf.

Jianhong Tu, Ju Fan, Nan Tang, Peng Wang, Guoliang Li, Xiaoyong Du, Xiaofeng Jia, and Song Gao. Unicorn: A unified multi-tasking model for supporting matching tasks in data integration. *Proceedings of the ACM on Management of Data*, 1(1):1–26, 2023. doi: 10.1145/3588938.

U.S. Department of Transportation, Bureau of Transportation Statistics. Airline on-time performance data. TranStats, 2026. URL https://www.transtats.bts.gov/DatabaseInfo.asp?QO_VQ=EFD. Monthly database, 1987–present; accessed 2026-08-30.

Runhui Wang, Yuliang Li, and Jin Wang. Sudowoodo: Contrastive self-supervised learning for multi-purpose data integration and preparation. In *Proceedings of IEEE ICDE*, pp. 1502–1515, 2023. doi: 10.1109/ICDE55515.2023.00391.

Sha Wang, Yuchen Li, Hanhua Xiao, Bing Tian Dai, Roy Ka-Wei Lee, Yanfei Dong, and Lambert Deng. Llmatch: A unified schema matching framework with large language models. In *Web and Big Data: APWeb-WAIM 2025*, volume 16116 of *Lecture Notes in Computer Science*, pp. 343–356. Springer, 2026. doi: 10.1007/978-981-95-5722-6_37.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems 33*, pp. 5776–5788, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115. URL https://arxiv.org/abs/2412.15115.

A Appendix

A.1 Resolver Algorithm

Algorithm 1 summarizes the constraint-aware global resolver used in Stage II.

Algorithm 1 Constraint-Aware Global Resolver

Require: Candidate graph $G = (V, E)$, compatibility masks Γ_{ij}, sampled penalty checks $\mathcal{Q}$, penalties β
Ensure: Feasible assignment (x^*, y^*)
1: Instantiate $x_{c,s}$, $y_{i,j}$, compatibility auxiliaries z_{ij,s_i,s_j}, and penalty slacks p_q
2: Add one-type-per-column constraints and linearization constraints for each $(s_i, s_j) \in \Gamma_{ij}$
3: Solve the relaxed program to obtain a fractional solution $(\tilde{x}, \tilde{y}, \tilde{z})$
4: Round type assignments column-by-column, keeping the highest feasible type
5: Sort candidate matches by fractional value and accept only those that preserve hard constraints
6: Decompose the graph into connected components and run deterministic local 2-opt improvement
7: **return** (x^*, y^*)

The algorithm makes explicit which steps drive consistency and which steps keep runtime manageable. The relaxation allows similarity scores, compatibility masks, and sampled penalties to interact before discrete commitments are made. Constraint-aware rounding prevents early local choices from forcing globally inconsistent matches. The connected-component improvement stage concentrates work where ambiguity remains instead of revisiting the entire lake, making the reported latency measurements interpretable at the level of the affected component.

A.2 Verification, Penalty, and Operational Details

A.2.1 Compatibility masks and linearization

For each candidate schema match (c_i, c_j), we precompute a sparse compatibility mask

$$\Gamma_{ij} = \{(s_i, s_j) \in \Gamma : (c_i, s_i) \in C_{\text{type}}, (c_j, s_j) \in C_{\text{type}}\}.$$

Only pairs in Γ_{ij} receive auxiliary variables z_{ij,s_i,s_j}, so the relaxation never instantiates the full $|\mathcal{S}|^2$ type-pair space. The linear constraints then encode the conjunction that c_i takes type s_i and c_j takes type s_j, while $y_{i,j}$ can activate only when at least one compatible pair is selected. This sparse construction is the practical reason the resolver stays tractable on the reported candidate graphs. It also aligns the compatibility logic used during optimization with the logic later used during verifier gating.

A.2.2 SCORE CALIBRATION AND COMPLETE RESOLVER CONSTRAINTS

The type scorer receives header-to-ontology-description cosine similarity, value-to-type-prototype cosine similarity, and the applicable string, numeric, or temporal sketch agreements. The match scorer receives header and value-embedding cosine similarities, MinHash/quantile/periodicity agreement, and posterior type compatibility. Unit-L_2 logistic scorers use the official train/validation/test partition where supplied and otherwise the same stratified 70/10/20 partition under each reported seed. They are fit on the training fold and Platt-calibrated on its validation fold; no test labels enter candidate retrieval, calibration, gating, or penalty selection. We use calibrated log-odds

$$w_{c,s}^{\mathrm{type}} = \log \frac{P_\theta(s \mid \phi(c))}{1 - P_\theta(s \mid \phi(c))}, \qquad w_{i,j}^{\mathrm{match}} = \log \frac{P_\psi(y_{i,j} = 1 \mid \phi(c_i), \phi(c_j))}{1 - P_\psi(y_{i,j} = 1 \mid \phi(c_i), \phi(c_j))}.$$

The null type $\perp$ is an abstention for columns lacking sufficient ontology evidence and participates in the same normalization. Because it carries no positive ontology information, it is compatible with every candidate type in Γ; its calibrated score prevents it from acting as a cost-free escape. For every column and candidate link, the relaxation includes

$$\sum_{s \in C(c) \cup \{\perp\}} x_{c,s} = 1, \quad z_{ij,s_i,s_j} \le x_{i,s_i}, x_{j,s_j}, \quad z_{ij,s_i,s_j} \ge x_{i,s_i} + x_{j,s_j} - 1,$$

$$y_{i,j} \le \sum_{(s_i,s_j) \in \Gamma_{ij}} z_{ij,s_i,s_j}, \qquad y_{i,j} = y_{j,i}.$$

For each candidate two-hop path $i-j-k$, we add $y_{i,j} + y_{j,k} - 1 \le y_{i,k}$ when $(i,k) \in E_{\mathrm{match}}$. If the closure edge is absent from the Stage I graph, the closed-world constraint $y_{i,j} + y_{j,k} \le 1$ prevents accepting both links. Transitivity is therefore explicit without creating a schema link outside the candidate graph. Each soft-rule slack is the mean of its linearized violation indicators, $p_q \ge |\mathcal{C}_q|^{-1} \sum_{r \in \mathcal{C}_q} \ell_r(x, y)$; this normalization keeps the fixed β_q values comparable across lakes with different check counts.

Rounding is deterministic given its evidence and initializer. Types are visited by decreasing fractional value, then links by decreasing fractional value; ties use larger absolute objective margin and then the stable edge identifier. Local refinement enumerates all feasible one-decision additions/deletions and two-decision swaps within a connected component in that same order, accepts the best positive objective change, and terminates after a complete pass with no improvement. It never introduces an edge outside the Stage I graph. The initial pass starts from the fixed candidate order. Verifier re-injection instead initializes the affected component from its current feasible assignment, applies the rule in Section 4.3, and runs the same deterministic rounding and local refinement once to produce the updated state.

A.2.3 VERIFIER PROMPT TEMPLATE

The verifier uses the following structured prompt.

```
You are a data integration expert. Determine if
these two columns represent the same attribute.

Column A:
- Header: "{header_a}"
- Sample values: {values_a}
- Inferred type: {type_a}

Column B:
- Header: "{header_b}"
- Sample values: {values_b}
- Inferred type: {type_b}

Ontology context:
- Type "{type_a}": {type_desc_a}
- Type "{type_b}": {type_desc_b}
```

```
- Compatible types: {compatible_types}

Output JSON with:
{
  "decision": "match" | "no_match" | "unsure",
  "confidence": 0.0-1.0,
  "rationale": "2-5 sentences explaining decision",
  "evidence": ["quoted patterns or values"]
}
```

This template is intentionally small so that prompt length remains stable across datasets. Fixing the evidence budget at 10 sampled values keeps token counts comparable across lakes and avoids hiding quality gains inside larger contexts. Injecting ontology snippets gives the verifier explicit compatibility information rather than asking it to infer type relations from language priors. The structured JSON output also makes auditing, caching, and failure analysis straightforward. Finally, the explicit `unsure` option lets the system fall back to the resolver instead of forcing a hallucinated binary decision.

A.2.4 Constraint penalty estimation and implementation details

Optimization sampling is capped at 10K stratified row-pair checks per FD dependency and 5K session-monotonicity cases for temporal penalties; unit penalties use currency and unit-pattern mismatches. The denominators reported in Table 7 are distinct evaluation counts: applicable test-fold check instances remaining after split assignment and validity filtering, rather than the raw optimization-sampling pools. The reported study keeps $\beta_{\text{FD}} = 2.0$, $\beta_{\text{unit}} = 3.0$, and $\beta_{\text{temp}} = 1.5$ fixed across datasets; the $\pm 20\%$ sensitivity band in Appendix Table 12 preserves the operating pattern. Headers use `all-MiniLM-L6-v2` and values use `roberta-base`, yielding a 1,152-dimensional neural representation; sketches enter the scorers as separate scalar features. HNSW retrieval uses `hnswlib` v0.7.0 with $M = 32$, $ef_{\text{construct}} = 200$, and $ef_{\text{search}} = 100$. NeuralMatch and Ditto use official splits when available and otherwise the shared 70/10/20 protocol under each reported seed, with 20 epochs, early stopping, batch size 64, sequence length 256, and AdamW at 2×10^{-5}. Software and hardware are Ubuntu 22.04, Python 3.11.7, PyTorch 2.2.2, Transformers 4.45.0, NumPy 1.26, Matplotlib 3.8, 2× Intel Xeon Gold 6248R CPUs, 512GB RAM, and 1× NVIDIA A100 80GB GPU.

A.2.5 Verifier Profile and Gating Sensitivity

Table 5: Verification profile by dataset. Each verified edge is one logical bundle; a cache miss makes three invocations, token columns report P50/P95 per executed invocation, and cache is the fraction of logical bundles served without an invocation.

Dataset	Verified	Cache	In toks	Out toks
WDC Products	2.7%	31%	420/690	150/240
WebTables	4.9%	24%	440/720	160/260
Flights+Weather	2.1%	41%	390/640	140/220
OSM Slices	2.9%	36%	410/670	150/240
MAG Profiles	3.2%	33%	405/665	145/235
Retail Lakehouse	4.2%	23%	470/780	175/290

Table 5 exposes model use at dataset granularity. Verification remains sparse even on the most ambiguous datasets, while input and output token quantiles stay comparable because every prompt fixes the evidence and ontology-context sizes. Cache reuse is highest when schema patterns recur across tables.

Table 6 shows that threshold tuning changes bundle volume much more than quality. The aggressive setting buys only a small Match-F1 gain over the default, while the conservative setting still exceeds the prompt-only baselines because the resolver has already done most of the structural work. The budget-capped row confirms that a hard per-component cap controls request concentration while preserving the main quality gain. This stable quality band is consistent with the interpretation that the result is not specific to one precise gate setting.

Table 6: Gating sensitivity averaged over the six public-core datasets and the CRM operational diagnostic.

SETTING	τ_I	τ_Δ	VERIFIED	MATCH F1
Conservative	0.07	0.12	2.2%	0.776
Balanced (default)	0.05	0.15	3.4%	0.785
Aggressive	0.03	0.18	4.8%	0.791
Budget-capped	0.03	0.18	3.9%	0.789

Table 6 summarizes sensitivity across the seven evaluated operating rows. Per-dataset request and token profiles appear in Table 5; the six-lake fixed-bundle comparison is reported separately in Table 10.

A.2.6 PUBLIC-CORE CONSISTENCY

Table 7: Per-dataset public-core consistency under the shared split, seed, candidate, and check protocol. Rate cells report LLM-Chain versus CanonWeave FD/unit/temporal percentages; intervals are paired 95% CIs for the reductions.

DATASET	LLM-CHAIN VS CANONWEAVE RATES	REDUCTION 95% CI: FD / UNIT / TEMPORAL	CHECKS: FD / UNIT / TEMPORAL
WDC Products	5.0/2.2/3.3 vs 1.8/1.0/1.5%	[2.7,3.7] / [0.9,1.5] / [1.4,2.2]	3,200 / 2,400 / 2,800
WebTables	5.6/2.7/3.9 vs 2.2/1.3/1.9%	[2.9,3.9] / [1.1,1.7] / [1.5,2.5]	3,500 / 2,500 / 3,100
Flights+Weather	4.3/2.0/2.9 vs 1.3/0.8/1.2%	[2.5,3.5] / [0.9,1.5] / [1.3,2.1]	2,700 / 1,900 / 2,300
OSM Slices	4.7/2.3/3.2 vs 1.4/0.9/1.3%	[2.8,3.8] / [1.1,1.7] / [1.5,2.3]	2,900 / 2,100 / 2,600
MAG Profiles	5.4/2.5/3.6 vs 1.9/1.1/1.7%	[3.0,4.0] / [1.1,1.7] / [1.5,2.3]	3,100 / 2,250 / 2,700
Retail Lakehouse	5.8/2.7/4.1 vs 1.6/0.9/1.4%	[3.6,4.8] / [1.5,2.1] / [2.2,3.2]	5,000 / 3,600 / 4,200
Unweighted mean	5.1/2.4/3.5 vs 1.7/1.0/1.5%	[3.0,3.8] / [1.1,1.7] / [1.6,2.4]	20,400 / 14,750 / 17,700

The reductions are positive for every violation class on every dataset. Retail exhibits the largest absolute improvement, while Flights retains the lowest final rates. This variation is consistent with the measured component-density pattern rather than a uniform threshold effect.

A.2.7 DETAILED SAME-GRAPH ORDERING CONTROL

Appendix Table 8 expands the total scheduling-policy intervention with marginal early-schedule Match intervals, consistency, verifier usage, latency, and raw counts; paired Match-effect intervals appear in Table 9. The candidate graph, resolver, verifier, token cap, gates, downstream ER stack, and hierarchical-bootstrap units are identical across schedules; the queried edge set and bundle count are policy outcomes. Early commitment uses more bundles on every dataset while producing a less consistent final state.

CanonWeave has the higher point estimate on every lake, and the paired Match-effect intervals in Table 9 exclude zero. Quality, consistency, and efficiency move together, with the largest bundle-volume and consistency penalties on Retail.

Table 8: Detailed same-graph scheduling control on six public-core datasets. CanonWeave Match/ER cells match Table 2; all early-verification cells use the same graph, resolver, verifier, token cap, gates, ER stack, and paired hierarchical-bootstrap protocol. Positive deltas favor CanonWeave.

DATASET	CANONWEAVE MATCH / ER	LLM-FIRST MATCH / ER; EARLY-MATCH 95% CI	FD / UNIT / TEMP.	VERIFIED / BUNDLES / P95 MIN	CANDIDATE / VERIFIED / GOLD / CHECKS
WDC Products	0.794 / 0.914	0.771 / 0.893; 0.763–0.778	2.4 / 1.3 / 2.0%	9.1% / 311 / 21.4m	3,420 / 311 / 1,146 / 8,400
WebTables	0.752 / 0.878	0.718 / 0.852; 0.711–0.725	3.1 / 1.8 / 2.6%	11.1% / 428 / 30.8m	3,868 / 428 / 1,324 / 9,100
Flights+Weather	0.831 / 0.921	0.792 / 0.897; 0.783–0.801	2.0 / 1.1 / 1.8%	7.6% / 176 / 14.6m	2,314 / 176 / 708 / 6,900
OSM Slices	0.812 / 0.918	0.784 / 0.894; 0.777–0.790	2.2 / 1.4 / 1.9%	8.3% / 241 / 19.7m	2,902 / 241 / 954 / 7,600
MAG Profiles	0.768 / 0.902	0.735 / 0.874; 0.728–0.742	2.7 / 1.5 / 2.1%	9.2% / 287 / 20.8m	3,116 / 287 / 1,098 / 8,050
Retail Lakehouse	0.781 / 0.893	0.727 / 0.861; 0.720–0.734	3.4 / 1.8 / 2.8%	13.1% / 493 / 33.5m	3,764 / 493 / 1,865 / 12,800
Public-core mean Δ	–	3.5 Match; 2.6 ER; 6/6 wins	0.9 / 0.5 / 0.7	+6.4 verified pts; 211 bundles	–

A.2.8 SAME-OUTPUT INJECTION REPLAY

The paired-policy experiment estimates the total effect of a scheduling rule, including the edges and bundles selected by that rule. Table X1 removes those mediators. For each seed, we freeze the late schedule's gate-selected edge set, execute each verifier bundle once, and cache the resulting JSON. One branch inserts those exact edge–output pairs into the unresolved initialization; the other first forms Z_0 and inserts them into affected components initialized from that assignment. Query identities, bundle count, outputs, confidence values, objective, feasibility rules, rounding, downstream computation, and deterministic tie-breaking are identical. The treatment is the initializer and resulting optimization path of the deterministic heuristic.

Appendix Table X1. Same-edge, same-bundle-count, same-output injection replay on six public-core datasets. Effects are late-minus-early gains for Match and ER F1 and early-minus-late reductions for the mean FD/unit/temporal violation rate; brackets give paired hierarchical 95% intervals.

DATASET	FIXED LOGICAL QUERIES	MATCH GAIN / ER GAIN / VIOLATION REDUCTION (95% CI)
WDC Products	92	+2.1 [+1.3,+2.9] / +1.5 [+0.9,+2.1] / +0.62 [+0.38,+0.86]
WebTables	190	+2.4 [+1.5,+3.3] / +1.8 [+1.1,+2.5] / +0.78 [+0.50,+1.06]
Flights+Weather	49	+1.6 [+0.8,+2.4] / +1.2 [+0.6,+1.8] / +0.47 [+0.25,+0.69]
OSM Slices	84	+0.8 [+0.1,+1.5] / +0.6 [+0.1,+1.1] / +0.31 [+0.12,+0.50]
MAG Profiles	100	+2.2 [+1.4,+3.0] / +1.7 [+1.0,+2.4] / +0.70 [+0.44,+0.96]
Retail Lakehouse	158	+3.0 [+2.0,+4.0] / +2.2 [+1.4,+3.0] / +1.02 [+0.72,+1.32]
Unweighted mean	–	+2.0 [+1.4,+2.6] / +1.5 [+1.0,+2.0] / +0.65 [+0.48,+0.82]; 6/6

Initialization from Z_0 has higher quality in all six replay comparisons. The 2.0 Match-F1 and 1.5 ER-F1 mean gains persist after removing query selection, output variation, and bundle count as explanations, while the 0.65 point mean-violation reduction quantifies path dependence under fixed evidence.

A.2.9 COMMITMENT INTERFERENCE

Table 9: Commitment interference under the same-graph control. Conflict measures pre-resolution commitments participating in a hard-compatibility or integrity conflict at first global coordination; overturn measures cases in which repair changes the committed edge itself rather than a neighbor or soft-rule slack; propagation counts downstream edge changes per overturn. Effect reports Match-F1 and mean-violation improvements from late verification with paired 95% CIs.

DATASET	CONFLICT: RATE / COUNT	OVERTURN: RATE / COUNT	PROPAGATION	EFFECT: MATCH / VIOLATION (95% CI)
WDC Products	24.4% / 76	64.5% / 49	2.3	+2.3 / +0.47 ([+1.5,+3.1] / [+0.32,+0.62])
WebTables	27.8% / 119	63.0% / 75	2.7	+3.4 / +0.70 ([+2.6,+4.2] / [+0.52,+0.88])
Flights+Weather	19.9% / 35	68.6% / 24	1.9	+3.9 / +0.53 ([+2.9,+4.9] / [+0.37,+0.69])
OSM Slices	20.3% / 49	59.2% / 29	1.8	+2.8 / +0.63 ([+2.0,+3.6] / [+0.46,+0.80])
MAG Profiles	31.7% / 91	65.9% / 60	3.1	+3.3 / +0.53 ([+2.5,+4.1] / [+0.38,+0.68])
Retail Lakehouse	42.0% / 207	66.2% / 137	3.8	+5.4 / +1.37 ([+4.5,+6.3] / [+1.12,+1.62])
Unweighted mean	27.7% / 96.2	64.6% / 62.3	2.6	+3.5 / +0.71

Interference increases with component density. Components with fewer than 10 candidate edges show 15.2% / 1.4 conflict/propagation; those with 10–40 edges show 28.9% / 2.6; and those above 40 edges reach 45.7% / 4.2. Retail and MAG contain the greatest share of dense components and exhibit the strongest coupled effect, whereas OSM's smaller components are consistent with its near-tie under the priority-greedy resolver.

A.2.10 PRIORITY-GREEDY RESOLVER INTERVENTION

Table X2 repeats the ordering intervention with the priority-greedy resolver defined in Section 4.2. Candidate graph, public-core splits, verifier, 4,096-token cap, downstream ER, and paired hierarchical-bootstrap protocol remain fixed.

Late verification wins 5/6 public-core datasets and improves mean Match, ER, consistency, and verifier traffic. OSM is a 0.1-point near-tie, aligning with its low commitment-interference rate in Table 9.

A.2.11 PROTOCOL-ADAPTED MODERN COMPARATORS

The "-class" suffix denotes a controlled public-core adapter rather than a transfer of numbers from each method's native benchmark. LLMATCH-class starts from the public v1.0.0 pipeline associated with Wang et al. (2026) and retains schema preparation, vector-similarity table selection, Rollup/Drilldown, and LLM column alignment. Schemora-class starts from the public pipeline associated with Gungor et al. (2025) and retains LLM metadata enrichment, vector-plus-BM25 retrieval, LLM table selection, and final candidate ranking. For both adapters, the native retrieval score ranks candidates inside the fixed Stage I public-core graph; raw headers, values, ontology descriptions, splits, and candidate universe are therefore shared, while each comparator keeps its published local-evidence modules.

Appendix Table X2. Matched priority-greedy ordering intervention on six public-core datasets. Each cell compares late residual verification with LLM-first → repair; the mean row is an unweighted macro-average.

DATASET	LATE VS EARLY MATCH / ER	EARLY-MATCH 95% CI	LATE VS EARLY FD / UNIT / TEMP.	LATE VS EARLY VERIF. / BUNDLES / P95 MIN	CAND. / LATE / EARLY / GOLD / CHECKS
WDC Products	0.787 / 0.909 vs 0.761 / 0.887	0.754–0.769	1.9 / 1.1 / 1.6% vs 2.5 / 1.4 / 2.0%	5.4% / 184 / 15.7m vs 7.2% / 247 / 18.9m	3,420 / 184 / 247 / 1,146 / 8,400
WebTables	0.747 / 0.871 vs 0.724 / 0.851	0.717–0.731	2.6 / 1.5 / 2.2% vs 3.0 / 1.7 / 2.5%	4.2% / 162 / 18.3m vs 5.8% / 223 / 21.4m	3,868 / 162 / 223 / 1,324 / 9,100
Flights+Weather	0.824 / 0.917 vs 0.795 / 0.896	0.787–0.803	1.5 / 0.9 / 1.4% vs 2.0 / 1.1 / 1.7%	3.2% / 73 / 11.6m vs 5.6% / 129 / 13.7m	2,314 / 73 / 129 / 708 / 6,900
OSM Slices	0.786 / 0.894 vs 0.787 / 0.895	0.780–0.794	1.7 / 1.1 / 1.5% vs 2.1 / 1.3 / 1.8%	3.2% / 92 / 13.8m vs 5.8% / 168 / 16.5m	2,902 / 92 / 168 / 954 / 7,600
MAG Profiles	0.761 / 0.895 vs 0.733 / 0.873	0.726–0.739	2.1 / 1.2 / 1.8% vs 2.6 / 1.5 / 2.1%	3.2% / 101 / 14.9m vs 5.8% / 181 / 18.2m	3,116 / 101 / 181 / 1,098 / 8,050
Retail Lakehouse	0.772 / 0.886 vs 0.748 / 0.867	0.741–0.756	1.9 / 1.1 / 1.7% vs 2.5 / 1.4 / 2.1%	5.0% / 189 / 19.8m vs 7.8% / 292 / 24.6m	3,764 / 189 / 292 / 1,865 / 12,800
Public-core mean Δ	2.2 Match; 1.7 ER; 5/6 wins	–	0.5 / 0.3 / 0.3	+2.3 verified pts; 73 bundles	–

We substitute the common GPT-4-Turbo verifier and 4,096-token cap, preserve each method's stage prompts, and serialize final match scores into one edge format consumed by the shared ER and consistency evaluators. Each adapter retains its local architecture, with its outputs flowing directly to the shared evaluators. Training-free modules remain training-free; any retrieval or threshold choice is selected on the validation fold from the defaults reported by the corresponding method, with Match F1 as the criterion and fewer verification bundles as the tie-breaker. At evaluation time, each method follows its native confidence ordering until it reaches exactly the per-dataset CanonWeave bundle count; every cache miss contains the same three self-consistency invocations, and each logical bundle is counted before cache lookup for all systems. Thus Table 10 compares adapter-level outcomes under common inputs, candidate support, verifier family, bundle count, and downstream measurement while retaining distinct architectures, prompts, preparation, enrichment, and ranking logic.

A.2.12 PUBLIC-CORE FIXED-BUDGET FRONTIER

Table 10 reports the six-public-core comparison with LLMATCH-class and Schemora-class adapters at the same per-dataset verification-bundle budget as CanonWeave.

Table 10: Public-core fixed-bundle comparison. Each comparator receives the exact per-dataset verification-bundle budget used by CanonWeave.

DATASET	SYSTEM	MATCH / ER	FD / UNIT / TEMP.	VERIFIED / BUNDLES / P95 MIN; DENOMINATORS
WDC Products	CanonWeave	0.794 / 0.914	1.8 / 1.0 / 1.5%	Reference bundle budget
	LLMATCH-class	0.773 / 0.899	3.1 / 1.8 / 2.5%	2.7% / 92 / 18.4m; 3,420 / 1,146 / 8,400

DATASET	SYSTEM	MATCH / ER	FD / UNIT / TEMP.	VERIFIED / BUNDLES / P95 MIN; DENOMINATORS
	Schemora-class	0.782 / 0.905	2.7 / 1.6 / 2.2%	2.7% / 92 / 17.6m
WebTables	CanonWeave	0.752 / 0.878	2.2 / 1.3 / 1.9%	Reference bundle budget
	LLMATCH-class	0.731 / 0.856	3.6 / 2.0 / 3.0%	4.9% / 190 / 27.1m; 3,868 / 1,324 / 9,100
	Schemora-class	0.742 / 0.864	3.2 / 1.8 / 2.7%	4.9% / 190 / 25.9m
Flights+Weather	CanonWeave	0.831 / 0.921	1.3 / 0.8 / 1.2%	Reference bundle budget
	LLMATCH-class	0.809 / 0.904	2.5 / 1.3 / 2.1%	2.1% / 49 / 11.9m; 2,314 / 708 / 6,900
	Schemora-class	0.814 / 0.908	2.3 / 1.2 / 1.9%	2.1% / 49 / 11.2m
OSM Slices	CanonWeave	0.812 / 0.918	1.4 / 0.9 / 1.3%	Reference bundle budget
	LLMATCH-class	0.797 / 0.904	2.8 / 1.6 / 2.2%	2.9% / 84 / 17.5m; 2,902 / 954 / 7,600
	Schemora-class	0.804 / 0.909	2.4 / 1.4 / 1.9%	2.9% / 84 / 16.7m
MAG Profiles	CanonWeave	0.768 / 0.902	1.9 / 1.1 / 1.7%	Reference bundle budget
	LLMATCH-class	0.756 / 0.888	3.2 / 1.7 / 2.5%	3.2% / 100 / 18.3m; 3,116 / 1,098 / 8,050
	Schemora-class	0.763 / 0.893	2.9 / 1.6 / 2.2%	3.2% / 100 / 17.6m
Retail Lakehouse	CanonWeave	0.781 / 0.893	1.6 / 0.9 / 1.4%	Reference bundle budget
	LLMATCH-class	0.745 / 0.869	3.8 / 2.1 / 3.1%	4.2% / 158 / 29.8m; 3,764 / 1,865 / 12,800
	Schemora-class	0.752 / 0.875	3.4 / 1.9 / 2.8%	4.2% / 158 / 28.7m

CanonWeave has the higher Match-F1 point estimate in all 12/12 controlled comparisons. Unweighted public-core advantages are +2.1 / +1.4 Match-F1 and +1.8 / +1.2 ER-F1 points over LLMATCH-class / Schemora-class, with paired 95% macro Match intervals of [+1.5,+2.8] / [+0.7,+2.0]. Mean FD/unit/temporal reductions are 1.5 / 0.8 / 1.1 ; 1.1 / 0.6 / 0.8 points.

A.2.13 CANDIDATE EXPOSURE DEGRADATION

Stage I determines the evidence available to every later decision. Table 11 reports exact gold-edge exposure counts for type and schema-match edges on all six public-core datasets. We then replay the full pipeline after retaining fractions $r \in \{1.00, 0.95, 0.90, 0.85, 0.80, 0.70\}$ of the undegraded graph's recovered exposure. Five randomized removals per level preserve candidate-class proportions and use the same resolver and verifier stack; paired hierarchical-bootstrap intervals summarize the public-core mean.

Table 11 makes candidate support an explicit operating coordinate. All public-core datasets recover most gold type and match edges, with useful variation in both ratios and counts. Retail and MAG expose the largest absolute edge sets, while Flights has the cleanest exposure profile, consistent with its simpler schema vocabulary and stronger type regularity. The normalized knee reports how much of each baseline recovered set can be removed before coordination loses its characteristic margin.

Table 11: Stage I exposure on six public-core datasets. Coverage is recovered/gold; the knee is the normalized fraction of baseline recovered exposure at the piecewise-linear Match-F1 bend.

DATASET	TYPE COVERAGE (RECOVERED / GOLD)	MATCH COVERAGE (RECOVERED / GOLD)	KNEE EXPOSURE / REMOVED EDGES
WDC Products	0.95 (203 / 214)	0.93 (1,067 / 1,146)	0.85 / 513
WebTables	0.92 (182 / 198)	0.90 (1,192 / 1,324)	0.84 / 619
Flights+Weather	0.97 (117 / 121)	0.96 (680 / 708)	0.80 / 463
OSM Slices	0.91 (160 / 176)	0.89 (849 / 954)	0.86 / 406
MAG Profiles	0.94 (219 / 233)	0.92 (1,010 / 1,098)	0.83 / 558
Retail	0.93 (248 / 267)	0.91 (1,697 / 1,865)	0.82 / 754
Lakehouse			

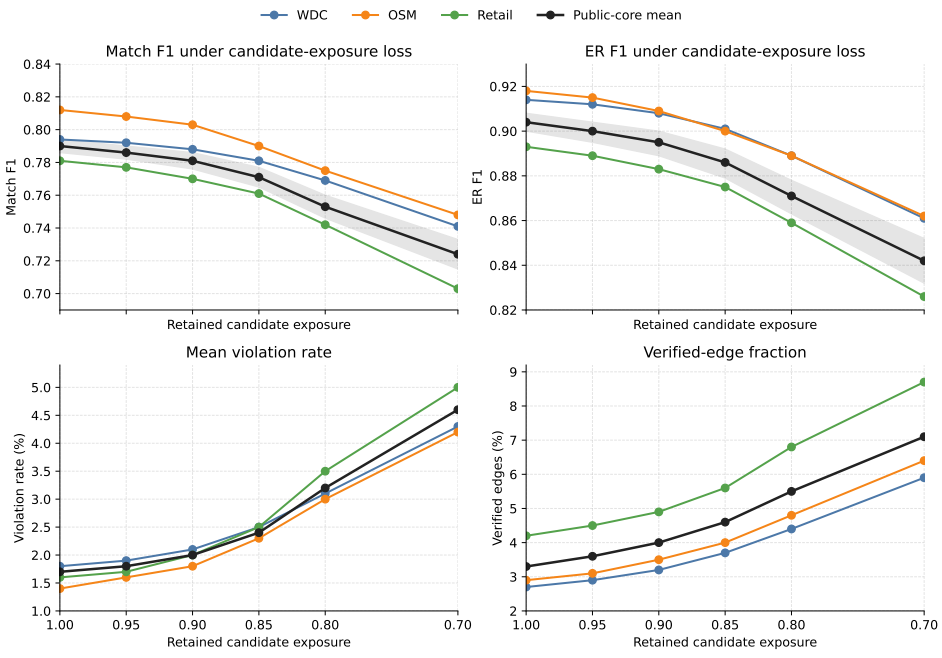


Figure 2: Candidate-exposure degradation. Curves retain $r \in \{1.00, 0.95, 0.90, 0.85, 0.80, 0.70\}$ of baseline recovered exposure; ribbons are hierarchical 95% intervals across five randomized degradations per level.

We estimate each knee at the maximum curvature of a piecewise-linear fit to mean Match F1 over the tested exposure levels; linear interpolation locates bends between grid points. Hierarchical intervals resample degradation seeds and the metric-specific structural units defined in Section 5.3, capturing both removal-pattern and evaluation uncertainty. The public-core mean retains most of its gain until 0.84 of baseline recovered exposure. WDC and OSM degrade smoothly as surviving candidates preserve more redundant constraint paths, while Retail bends earlier under its denser alias graph.

A.2.14 Penalty, Ontology, and Verifier Robustness

We also test whether the measured operating point survives mild regime shift on two selected medium-scale datasets (WDC Products, OSM Slices) and one larger lake (Retail Lakehouse). Table 12 keeps the main protocol fixed while applying three perturbations: a mild ontology degradation that corrupts 10% of compatibility edges, a $\pm 20\%$ sweep around the fixed β vector, and a second verifier family instantiated with Qwen2.5-72B-Instruct under the same gating policy.

Table 12: Robustness under 10% ontology-edge corruption, a $\pm 20\%$ penalty-weight sweep, and Qwen2.5-72B-Instruct on three selected public-core datasets.

Dataset	10% ontology-edge corruption: Match; FD / Unit / Temp.; verif. / P95 min	β-band ($\pm 20\%$): Match; FD / Unit / Temp.; verif. / P95 min	Qwen2.5-72B-Instruct: Match; FD / Unit / Temp.; verif. / P95 min
WDC Products	0.783; 2.1 / 1.2 / 1.8%; 3.4% / 11.8m	0.789–0.797; 1.8 / 1.0 / 1.5% – 2.0 / 1.1 / 1.7%; 2.5%–3.0% / 10.9–11.4m	0.788; 1.9 / 1.1 / 1.6%; 3.1% / 17.6m
OSM Slices	0.800; 1.8 / 1.1 / 1.6%; 3.7% / 13.2m	0.807–0.815; 1.4 / 0.9 / 1.3% – 1.7 / 1.0 / 1.5%; 2.6%–3.3% / 12.4–12.8m	0.803; 1.6 / 1.0 / 1.5%; 3.0% / 19.3m
Retail Lakehouse	0.752; 2.8 / 1.6 / 2.4%; 5.1% / 21.6m	0.771–0.786; 1.8 / 1.0 / 1.5% – 2.1 / 1.2 / 1.8%; 3.8%–4.6% / 20.1–20.9m	0.768; 2.0 / 1.1 / 1.7%; 4.4% / 28.7m

The robustness results establish a coherent qualitative ordering across the tested shifts. Mild ontology damage and verifier-family swaps reduce quality smoothly, and modest β perturbations preserve the same sparse-verification regime. WDC and OSM remain relatively stable under the three perturbations. Retail is the most sensitive case under both ontology corruption and the slower verifier family, consistent with its larger, denser alias graph; request volume and latency move before quality changes sharply.

A.2.15 Symbolic-Substrate Degradation

Table 13 replays all six public-core datasets under 1.00 / 0.90 / 0.75 / 0.60 retained structure; five degradations per level. At each level we retain a stratified random subset of ontology-compatibility edges and FD/unit/temporal rules within their relation classes, without consulting gold labels; degradation seeds independently sample the removed structure. The four levels retain 42,680/1,284; 38,412/1,156; 32,010/963; 25,608/770 compatibility edges/rules. Each cell compares late residual verification against the same early schedule as the main intervention.

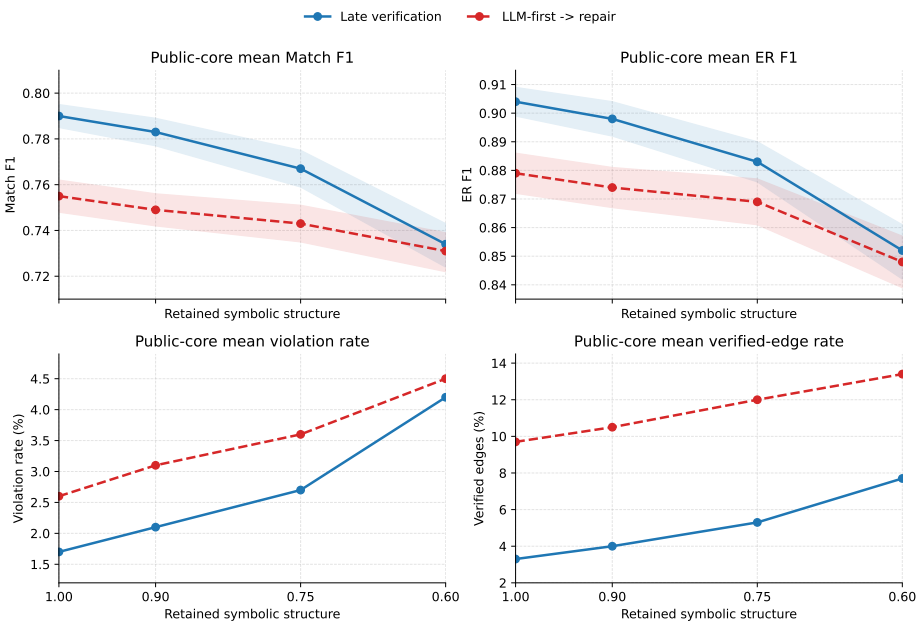
Table 13: Symbolic-substrate stress test on six public-core datasets. Every level compares late and early verification over five stratified degradations.

Dataset	Retained	Late vs early Match / ER	Late vs early FD / Unit / Temp.	Late vs early verified / checks
WDC Products	1.00	0.794 / 0.914 vs 0.771 / 0.893	1.8 / 1.0 / 1.5% vs 2.4 / 1.3 / 2.0%	2.7% / 5x8,400 vs 9.1% / 5x8,400
WDC Products	0.90	0.789 / 0.909 vs 0.766 / 0.889	2.1 / 1.1 / 1.7% vs 2.7 / 1.5 / 2.2%	3.4% / 5x8,400 vs 10.0% / 5x8,400

DATASET	RETAINED	LATE VS EARLY MATCH / ER	LATE VS EARLY FD / UNIT / TEMP.	LATE VS EARLY VERIFIED / CHECKS
WDC Products	0.75	0.775 / 0.896 vs 0.760 / 0.882	2.8 / 1.5 / 2.3% vs 3.4 / 1.9 / 2.8%	4.9% / 5x8,400 vs 11.9% / 5x8,400
WDC Products	0.60	0.744 / 0.868 vs 0.739 / 0.862	4.1 / 2.2 / 3.4% vs 4.6 / 2.5 / 3.8%	7.3% / 5x8,400 vs 13.2% / 5x8,400
WebTables	1.00	0.752 / 0.878 vs 0.718 / 0.852	2.2 / 1.3 / 1.9% vs 3.1 / 1.8 / 2.6%	4.9% / 5x9,100 vs 11.1% / 5x9,100
WebTables	0.90	0.744 / 0.871 vs 0.710 / 0.844	2.6 / 1.5 / 2.2% vs 3.5 / 2.0 / 2.9%	5.5% / 5x9,100 vs 12.0% / 5x9,100
WebTables	0.75	0.729 / 0.858 vs 0.708 / 0.843	3.4 / 1.9 / 2.8% vs 4.0 / 2.3 / 3.4%	6.9% / 5x9,100 vs 13.4% / 5x9,100
WebTables	0.60	0.701 / 0.833 vs 0.697 / 0.829	4.8 / 2.7 / 3.9% vs 5.2 / 2.9 / 4.3%	8.8% / 5x9,100 vs 14.9% / 5x9,100
Flights+Weather	1.00	0.831 / 0.921 vs 0.792 / 0.897	1.3 / 0.8 / 1.2% vs 2.0 / 1.1 / 1.8%	2.1% / 5x6,900 vs 7.6% / 5x6,900
Flights+Weather	0.90	0.826 / 0.916 vs 0.788 / 0.893	1.5 / 0.9 / 1.4% vs 2.2 / 1.2 / 1.9%	2.6% / 5x6,900 vs 8.2% / 5x6,900
Flights+Weather	0.75	0.813 / 0.906 vs 0.784 / 0.890	2.1 / 1.2 / 1.8% vs 2.7 / 1.5 / 2.3%	3.8% / 5x6,900 vs 9.5% / 5x6,900
Flights+Weather	0.60	0.781 / 0.878 vs 0.776 / 0.873	3.4 / 1.9 / 2.9% vs 3.8 / 2.2 / 3.2%	5.9% / 5x6,900 vs 11.0% / 5x6,900
OSM Slices	1.00	0.812 / 0.918 vs 0.784 / 0.894	1.4 / 0.9 / 1.3% vs 2.2 / 1.4 / 1.9%	2.9% / 5x7,600 vs 8.3% / 5x7,600
OSM Slices	0.90	0.806 / 0.911 vs 0.779 / 0.889	1.7 / 1.0 / 1.5% vs 2.5 / 1.6 / 2.2%	3.5% / 5x7,600 vs 9.0% / 5x7,600
OSM Slices	0.75	0.791 / 0.897 vs 0.778 / 0.885	2.4 / 1.4 / 2.1% vs 3.0 / 1.9 / 2.7%	4.7% / 5x7,600 vs 10.4% / 5x7,600
OSM Slices	0.60	0.756 / 0.862 vs 0.754 / 0.860	3.9 / 2.2 / 3.4% vs 4.2 / 2.5 / 3.7%	7.1% / 5x7,600 vs 11.8% / 5x7,600
MAG Profiles	1.00	0.768 / 0.902 vs 0.735 / 0.874	1.9 / 1.1 / 1.7% vs 2.7 / 1.5 / 2.1%	3.2% / 5x8,050 vs 9.2% / 5x8,050
MAG Profiles	0.90	0.761 / 0.895 vs 0.728 / 0.867	2.2 / 1.3 / 1.9% vs 3.0 / 1.7 / 2.4%	3.9% / 5x8,050 vs 9.9% / 5x8,050
MAG Profiles	0.75	0.744 / 0.880 vs 0.726 / 0.864	2.9 / 1.6 / 2.5% vs 3.5 / 2.0 / 3.0%	5.2% / 5x8,050 vs 11.3% / 5x8,050
MAG Profiles	0.60	0.714 / 0.849 vs 0.712 / 0.846	4.3 / 2.4 / 3.8% vs 4.7 / 2.7 / 4.1%	7.8% / 5x8,050 vs 12.9% / 5x8,050
Retail Lakehouse	1.00	0.781 / 0.893 vs 0.727 / 0.861	1.6 / 0.9 / 1.4% vs 3.4 / 1.8 / 2.8%	4.2% / 5x12,800 vs 13.1% / 5x12,800
Retail Lakehouse	0.90	0.773 / 0.885 vs 0.724 / 0.857	1.9 / 1.1 / 1.7% vs 3.7 / 2.0 / 3.1%	4.9% / 5x12,800 vs 13.9% / 5x12,800
Retail Lakehouse	0.75	0.752 / 0.863 vs 0.726 / 0.851	2.7 / 1.5 / 2.4% vs 4.2 / 2.4 / 3.7%	6.5% / 5x12,800 vs 15.4% / 5x12,800
Retail Lakehouse	0.60	0.709 / 0.821 vs 0.708 / 0.820	4.6 / 2.7 / 4.1% vs 4.8 / 2.9 / 4.3%	9.4% / 5x12,800 vs 16.8% / 5x12,800

Table 13 maps ordering behavior over materially weaker symbolic states. The high-retention rows reproduce the main pattern; intermediate rows contract gradually; and the lowest-retention rows locate the near-zero margin, especially on the largest lakes. The resulting curve shows directly how compatibility paths and usable rules determine the advantage available from global coordination.

We locate the half-margin point by linearly interpolating unrounded public-core Match-F1 differences between adjacent retention levels. Hierarchical intervals resample datasets, degradation seeds, and the metric-specific structural units defined in Section 5.3. The advantage halves near 0.72 retained



Appendix Figure X1. Verifier-placement phase diagram over symbolic support. The state-conditioned advantage halves near 0.72 retained structure and preserves its direction through 0.60 retained structure.

structure, with its direction preserved through the lowest evaluated 0.60 retained structure. Retail and MAG consume verifier budget fastest as compatibility paths disappear.

A.2.16 All-Six Qwen Ordering Intervention

Table X3 repeats the same-graph intervention with Qwen2.5-72B-Instruct. Candidate graph, resolver, prompt schema, token cap, gate, splits, seeds, downstream ER, and bootstrap protocol remain fixed; the state-conditioned policy changes prompt context, selected edges, bundle volume, and initializer across the early and late schedules. The main public-core GPT-4-Turbo rows anchor the model-swap effect at the late-verification operating point.

Qwen late verification beats Qwen early commitment on all six datasets, with the largest margin on Retail. At the late operating point, replacing GPT-4-Turbo with Qwen changes public-core mean Match F1 by 1.0 point and increases P95 latency. Ordering is stable across verifier families even as model choice moves the quality–latency point.

A.3 Extended Evidence and Failure Analysis

This subsection collects per-dataset ablations, uncertainty, and illustrative mechanisms.

A.3.1 Per-Dataset Ablations

Table 14 confirms the aggregate pattern on every dataset. Ontology compatibility contributes the largest gains, especially on the CRM and Retail slices where aliasing and rule density are high. Soft penalties consistently reduce the FD violations shown here, aligning canonical-state quality with integrity. Selective verification then improves Match and ER after global structure has resolved most decisions, demonstrating the value of targeted learned support. The dataset-level consistency of these effects reinforces the mechanism established by the ordering interventions.

A.3.2 Confidence Intervals

All reported intervals preserve the lake-level dependence induced by joint resolution. Each of 10K paired hierarchical replicates first resamples the five seed runs, then resamples structural units

Appendix Table X3. Qwen2.5-72B ordering intervention under the shared six-public-core protocol.

Dataset	Qwen late vs early M / ER (early-M 95% CI)	Late vs early FD / unit / temporal	Late vs early verified / bundles / P95 min	Candidate / late / early / gold / checks
WDC Products	0.788 / 0.910 vs 0.766 / 0.889 (0.758,0.774)	1.9/1.1/1.6 vs 2.5/1.4/2.0%	3.1%/107/17.6m vs 8.0%/274/22.8m	3,420 / 107 / 274 / 1,146 / 8,400
WebTables	0.741 / 0.869 vs 0.713 / 0.847 (0.705,0.721)	2.5/1.5/2.1 vs 3.3/1.9/2.7%	4.7%/182/24.9m vs 10.2%/394/31.5m	3,868 / 182 / 394 / 1,324 / 9,100
Flights+ Weather	0.824 / 0.916 vs 0.793 / 0.895 (0.784,0.802)	1.5/0.9/1.4 vs 2.1/1.2/1.8%	2.6%/60/12.1m vs 7.1%/164/17.2m	2,314 / 60 / 164 / 708 / 6,900
OSM Slices	0.803 / 0.910 vs 0.781 / 0.891 (0.773,0.789)	1.6/1.0/1.5 vs 2.2/1.4/2.0%	3.0%/87/19.3m vs 7.7%/224/25.0m	2,902 / 87 / 224 / 954 / 7,600
MAG Profiles	0.756 / 0.894 vs 0.730 / 0.870 (0.722,0.738)	2.2/1.3/1.9 vs 2.8/1.6/2.3%	3.4%/106/18.8m vs 8.6%/268/26.1m	3,116 / 106 / 268 / 1,098 / 8,050
Retail Lakehouse	0.768 / 0.883 vs 0.724 / 0.854 (0.716,0.732)	2.0/1.1/1.7 vs 3.5/1.9/2.9%	4.4%/166/28.7m vs 12.4%/467/36.8m	3,764 / 166 / 467 / 1,865 / 12,800
Mean late advantage	+2.9 Match / +2.3 ER; 6/6 wins	0.8 / 0.4 / 0.6	5.5 pts / 181 bundles	–

Table 14: Per-dataset component ablations.

DATASET	ABLATION	TYPE	MATCH	ER	FD VIOL.
WDC	Full	0.851	0.794	0.914	1.8%
	-Ontology	0.806	0.721	0.868	4.2%
	-Soft	0.824	0.752	0.884	2.4%
	-LLM	0.821	0.748	0.878	2.1%
WebTables	Full	0.823	0.752	0.878	2.2%
	-Ontology	0.791	0.704	0.854	5.7%
	-Soft	0.803	0.720	0.856	3.1%
	-LLM	0.806	0.711	0.848	2.9%
Flights	Full	0.904	0.831	0.921	1.3%
	-Ontology	0.881	0.782	0.892	4.4%
	-Soft	0.890	0.806	0.912	2.1%
	-LLM	0.884	0.794	0.904	1.9%
SaaS CRM (diag.)	Full	0.834	0.758	0.874	1.5%
	-Ontology	0.780	0.689	0.827	5.1%
	-Soft	0.806	0.730	0.856	2.8%
	-LLM	0.794	0.708	0.838	2.2%
OSM	Full	0.876	0.812	0.918	1.4%
	-Ontology	0.827	0.739	0.880	5.3%
	-Soft	0.842	0.772	0.896	2.6%
	-LLM	0.844	0.766	0.888	2.2%
MAG	Full	0.842	0.768	0.902	1.9%
	-Ontology	0.804	0.719	0.870	4.5%
	-Soft	0.820	0.738	0.886	2.7%
	-LLM	0.818	0.731	0.878	2.4%
Retail	Full	0.847	0.781	0.893	1.6%
	-Ontology	0.795	0.714	0.856	4.9%
	-Soft	0.832	0.718	0.856	2.6%
	-LLM	0.800	0.736	0.863	2.1%

within each selected run: tables for Type F1, connected candidate components for Match F1 and FD/unit/temporal checks, and entity blocks for ER F1. Both methods contribute predictions for the same sampled units before their difference is computed. Dataset-level effect intervals use that paired difference distribution; public-core summaries recompute the unweighted six-dataset mean in every replicate. Neither edges from one coupled component nor records from one entity block are treated as independent observations.

Table 15: Marginal Match-F1 confidence intervals under the common LLM-Chain comparison.

DATASET	LLM-CHAIN MATCH F1 (95% CI)	CANONWEAVE MATCH F1 (95% CI)
WDC Products	0.738 (0.732, 0.744)	0.794 (0.789, 0.799)
WebTables	0.691 (0.683, 0.699)	0.752 (0.746, 0.758)
Flights+Weather	0.768 (0.757, 0.779)	0.831 (0.822, 0.840)
SaaS CRM (diag.)	0.648 (0.640, 0.656)	0.758 (0.751, 0.765)
OSM Slices	0.761 (0.754, 0.768)	0.812 (0.806, 0.818)
MAG Profiles	0.702 (0.694, 0.710)	0.768 (0.762, 0.774)
Retail Lakehouse	0.684 (0.676, 0.692)	0.781 (0.774, 0.788)

Table 15 shows nonoverlapping marginal intervals for CanonWeave and LLM-Chain. Flights has the widest interval because it contains fewer match pairs, while larger lakes combine narrower intervals with strong gains. Paired effect intervals for the scheduling intervention are reported separately in Tables X1 and 9.

A.3.3 FAILURE CASES

Table 16: Illustrative baseline errors and the CanonWeave decisions that correct them.

CASE	COLUMNS	BASELINE ERROR	CANONWEAVE ACTION	DECISIVE SIGNAL
Unit mismatch	`price_usd`, `price_eur`	Local similarity merges comparable numeric ranges	Rejects the schema link	Currency-unit incompatibility
Type confusion	`city`, `product_name`	Local type scoring assigns broad string types; the independent prompt then links them	Assigns distinct types and rejects the link	Ontology compatibility
Lexical alias	`order_ts`, `created_at`	Embedding matcher misses the correspondence	Verifies and accepts the residual link	Type context, LLM evidence, and FD support
Verifier instability	`cust_id`, `customer_id`	Prompt variants disagree on the local decision	Falls back to the resolver, which accepts the link	Self-consistency and FD support

Table 16 complements the aggregate metrics with concrete failure modes. The unit-mismatch example shows why numeric similarity alone is unreliable when currencies differ. The type-confusion example shows that lexical similarity is insufficient when two attributes share a broad surface form but belong to incompatible ontology types. The alias-resolution example shows where selective verification contributes genuinely new evidence after the resolver has narrowed the search space. The LLM-instability example shows why self-consistency and resolver fallback are necessary even when the verifier is strong.

In the unit-mismatch case, `price_usd` and `price_eur` have high embedding similarity because both columns contain comparable numeric ranges. A purely local matcher therefore overestimates the chance that they represent the same attribute. CanonWeave rejects the edge because currency symbols trigger the unit penalty before the match is committed. The case matters because these false

positives directly corrupt canonical schemas and downstream joins. It illustrates why consistency metrics can move with quality in the main paper.

In the type-confusion case, `city` and `product_name` both contain capitalized strings and are therefore difficult to separate with local text features alone. The local type scorer can assign broad string types, after which independent prompt matching proposes an incorrect link. The ontology mask prevents this because CITY and PRODUCT are incompatible in the partial ontology. The example shows that symbolic structure does real work even before any verifier call is made. It also clarifies why removing ontology compatibility causes the largest ablation drop.

In the alias-resolution case, `order_ts` and `created_at` have weak lexical similarity despite representing the same timestamp attribute. Neural baselines miss the edge because the names diverge and the local value summaries are short. The verifier succeeds only after the resolver has already narrowed the ambiguity set and supplied explicit type context. The supporting FD pattern order_id $\rightarrow$ timestamp then makes the verified match globally coherent. This case illustrates where sparse verification earns its keep.

In the instability case, `cust_id` and `customer_id` are clearly the same attribute, but prompt formatting changes still cause occasional verifier disagreement. Self-consistency catches the disagreement and the system falls back to the resolver when no stable majority emerges. This matters because a single unstable local answer should not rewrite the lake-level state. The case also explains why exhaustive verification can amplify unstable local evidence. More generally, it demonstrates the value of routing verifier judgments through a consistent global program.

Together, the failure cases and controlled interventions identify the same mechanism at two scales: globally coupled mistakes arise when local evidence becomes binding too early, while CanonWeave's resolved-state arbitration confines learned judgments to residual decisions that improve a feasible canonical schema.