
Non-Axiomatic Reasoning for Longitudinal Identity Resolution: NARS on ICE-ID and Standard ER Benchmarks

Gonçalo Hora de Carvalho*
IIIM, Iceland
goncalo@iiim.is

Lazar S. Popov
IIIM, Iceland

Sander Kaatee
IIIM, Iceland

Kristinn R. Thórisson
Full Research Professor, Department of Computer Science
Reykjavik University

Tangrui Li
Temple University

Pétur Húni Björnsson
Department of Nordic Studies and Linguistics
University of Copenhagen

Jilles S. Dibangoye
Associate Professor, Machine Learning Group, Department of Artificial Intelligence
Bernoulli Institute, University of Groningen

Abstract

We evaluate OpenNARS-for-Applications (ONA) for longitudinal identity resolution on Icelandic historical census data spanning 220 years (1703–1920), while benchmarking it on standard entity resolution datasets (Abt–Buy, Amazon–Google, DBLP–ACM, DBLP–Scholar) [? 3]. NARS is a general-purpose AI framework designed to reason with limited knowledge and resources; its core is Non-Axiomatic Logic (NAL), a term-based logic. Our experiments show that this evidence-based symbolic matcher can serve as a transparent baseline alongside classical probabilistic linkage and modern neural matchers. We report pairwise metrics, calibration sensitivity, and a deployment-faithful candidate-graph evaluation on ICE-ID, and compare against strong baselines across standard ER datasets.

1 Introduction

Linking historical census records is an important step in research on social mobility, demographic change, migration, and epidemiology, yet it remains arduous because names mutate, fields are missing, and administrative borders shift over time [6?]. While supervised matchers trained on labeled pairs improved accuracy for nineteenth-century U.S. censuses [?], most census-specific benchmarks still oversimplify the problem: they cover only short time ranges (often a single decade), omit kinship structure, and treat geography as flat text rather than a hierarchy [5].

Benchmark design strongly shapes progress in entity resolution and record linkage. We release **ICE-ID** to fill a missing setting: long-horizon longitudinal person matching with temporal drift, hierarchical geography, and expert-curated identity clusters. Spanning 16 Icelandic census waves

*Corresponding author: goncalo@iiim.is

(1703–1920) and covering more than 220 years, ICE-ID contains 984 028 rows and 226,864 expert-curated person clusters.

We formulate two tasks:

1. **Intra-census linkage:** identify the same individual within a single census.
2. **Cross-census linkage:** identify the same individual across successive censuses despite spelling drift and age progression.

Train/validation/test splits follow a strictly temporal protocol: rows up to 1870 for training, 1871–1890 for validation, and 1891–1920 for held-out testing, mirroring real archival workflows. Evaluation combines pairwise metrics (precision, recall, F_1 , ROC-AUC) with clustering quality (Adjusted Rand Index, ARI).

We benchmark several model families:

- *Classical probabilistic linkage:* Fellegi–Sunter and deterministic rule-based matchers;
- *Symbolic reasoning:* a Non-Axiomatic Reasoning System (NARS) that learns identity rules from streaming examples;
- *Neural matchers:* Ditto [?], HierGAT [?], and zero-shot/LLM-based methods (AnyMatch [?], MatchGPT).

Our results show that NARS achieves competitive pairwise accuracy ($F_1=0.994$ on ICE-ID) using only symbolic judgment-based scoring, without deep representation learning or GPU resources. On standard ER benchmarks (DBLP–ACM, DBLP–Scholar), NARS matches or exceeds reported state-of-the-art F_1 , while remaining below SOTA on product-matching datasets (Abt–Buy, Amazon–Google). These findings suggest that symbolic reasoning can serve as a strong, transparent baseline for entity resolution, and motivate *hybrid* pipelines that fuse symbolic and neural evidence.

2 Methods

This section describes the ICE-ID benchmark, the NARS matcher, baselines, and evaluation protocol.

2.1 ICE-ID Benchmark

We evaluate on ICE-ID [?], a benchmark of 984,028 Icelandic census records (1703–1920) with 226,864 expert-labeled person clusters. We use the canonical temporal splits (up to 1870 train / 1871–1890 validation / 1891–1920 test) and report pairwise metrics (Precision, Recall, F_1 , AUC), clustering metrics (ARI, B^3), and ranking metrics ($P@k$, $R@k$). Full dataset documentation is in the companion data paper.

2.1.1 Data Preparation

Records are loaded from `people.csv` with string columns normalized to lowercase and numeric fields (`id`, `heimild`, `birthyear`, `person`, and geographic/kinship IDs) coerced to integers. Missing IDs are dropped; missing name fields are filled with empty strings. NARS and the symbolic matchers operate directly on these record pairs; the ML ensemble baseline uses pairwise TF-IDF features over concatenated record text. All ICE-ID experiments under the shared protocol use the same raw record representation and temporal splits keyed by `heimild`.

2.1.2 Pair Construction and Sampling

Positive pairs are drawn from expert-labeled clusters: two records share a positive label if they share the same person (or equivalent) ID. For *cross-wave* linkage, the benchmark supports pair generation across distinct census-year windows (e.g., source and target year ranges) via the same cluster semantics. Negatives are formed by sampling pairs whose records belong to different clusters. We use a 2:1 negative-to-positive ratio when building the labeled set within each split. For ICE-ID, pair generation and any subsampling are performed *inside* the fixed temporal splits (train: up to 1870, validation: 1871–1890, test: 1891–1920); no random train/validation/test split is applied across

years. If the number of candidate pairs in a split exceeds 50,000, we uniformly subsample that split to the cap so that NARS and the ML baselines are evaluated on the same bounded pair set. The implementation uses a configurable cap of 50,000 pairs and does not impose a same-census-year constraint on negatives; within-wave evaluation can be obtained by restricting the record pool to a single heimild in a given run.

2.2 ML Ensemble Baseline

As a baseline, we compare against an ensemble of four tree-based classifiers (XGBoost, LightGBM, CatBoost, Random Forest) averaging match probabilities. The ensemble is trained on pairwise text-derived and structured comparison features computed from the same raw record pairs used by NARS. Model selection uses 10-fold cross-validation within the training split; final evaluation uses the same held-out validation and test splits as NARS. Full details are provided in the ICE-ID data paper.

2.3 OpenNARS-for-Applications

Implementation: We use OpenNARS-for-Applications (ONA) as the underlying NARS engine. ONA defines the judgment vocabulary and evidence framework; record pairs are converted into Narsese statements encoding attribute agreements and disagreements. The reported experiments use the LLR scoring instantiation described in the “Practical scoring variant” paragraph below.

Algorithm Overview: Figure 1 summarizes the procedure.

```

Inputs:  labeled training pairs (r1, r2, y), pool size P, patterns used n
1) For each training pair:
   J = preprocess(r1, r2)                                     (atomic judgments)
   tv = TruthValue(f=y, c=0.9)
   add/update (J → tv) in PatternPool (evict if > P)
2) Calibrate a threshold  $\tau$  on held-out labeled pairs
   (ICE-ID: median midpoint; classic ER: validation  $F_1$ )
3) For each test pair:
   Jq = preprocess(r1, r2)
   select n reference patterns from pool (top/bottom expectation)
   score = revise evidence from overlaps between Jq and selected patterns
   predict match if score  $\geq \tau$ 

```

Figure 1: Conceptual NARS entity resolution procedure (pseudocode). Reported experiments use the LLR scoring variant described below.

This pseudocode illustrates the conceptual pattern-pool formulation for intuition; the reported implementation replaces Step 3’s revision-based scoring with the LLR scorer described below.

Complexity: Conceptual pattern-pool scoring is $O(N \cdot n \cdot |J|)$ where N = pairs, n = reference patterns, $|J|$ = judgments per pair; the LLR variant reduces to $O(N \cdot |J|)$ since each judgment maps to a precomputed log-ratio.

The LLR scoring pipeline used in all reported experiments proceeds as follows:

1. **Judgment Generation.** A preprocessing function (`preprocess_iceid`) converts each record pair (r_1, r_2) into a set of atomic judgments (e.g., `same_nafn_norm`, `birthyear_close`) based on attribute comparisons for name, birth year, sex, census year (heimild), and location.
2. **Evidence Accumulation.** During training, each labeled pair increments per-judgment positive or negative counts (with Laplace +1 smoothing). The accumulated counts yield a log-likelihood ratio $LLR_j = \log \frac{P(j|match)}{P(j|non-match)}$ for every judgment type j .
3. **Pair Scoring.** For a new pair, the generated judgments are looked up in the LLR table, summed with a base-rate log-prior, and passed through a sigmoid to produce a match probability.

2.3.1 Conceptual NARS Formulation (Background)

The following pattern, pattern-pool, and revision rules describe the conceptual NARS formulation that motivates the design. They establish the correspondence with Non-Axiomatic Logic (see Appendix A.4) but are *not* used in the reported experiments, which rely on the LLR scoring pipeline above.

Pattern. A pair of census records yields an atomic pattern: the unordered set of all generated judgments, annotated with truth-value $(1, 0.9)$ if the records belong to the same individual or $(0, 0.9)$ otherwise.

Pattern Pool. The pattern pool contains patterns, sorted in ascending order of the expectation of the truth-value.

Inference Rule. Given two patterns p_1 and p_2 containing judgment sets j_1 and j_2 with truth-values t_1 and t_2 , the proposed inference rule derives three new patterns:

1. The judgment set of the new pattern is $j_1 \setminus j_2$, and the truth-value is t_1 .
2. The judgment set of the new pattern is $j_2 \setminus j_1$, and the truth-value is t_2 .
3. The judgment set of the new pattern is $j_1 \cap j_2$, and the truth-value is $\text{revise}(t_1, t_2)$.

These rules ensure that each derived truth-value is inherited from parent evidence rather than fabricated. In case 1, evidence contributing to t_1 also supports the new sub-pattern; its truth-value may later be revised with independently obtained patterns. Case 2 is symmetric.

In case 3, the intersection pattern is a sub-pattern of both parents, so both t_1 and t_2 contribute evidence and revision merges them. To avoid double-counting, only patterns derived from independent evidence sources may participate in a single revision step.

Learning as Recognition. In the conceptual formulation, the system constructs a query pattern and measures its overlap with existing patterns in the pool. The degree of overlap determines the match score. During this process, new patterns may be added to the pool and existing truth-values may be updated.

In the conceptual formulation, the final expectation score can be thresholded to obtain a binary match decision. In the reported implementation, threshold calibration is performed in the deployed LLR pipeline described below.

Conceptual truth-value scoring (not used in reported experiments). At scoring time, up to n reference patterns are retrieved from the pool—half with the highest expectation (likely matches) and half with the lowest (likely non-matches)—to provide balanced evidence. For each reference pattern, a pairwise truth-value is computed: frequency f equals the size of the judgment intersection divided by the size of the longer pattern, and confidence c is inherited from the longer (more specific) pattern. These pairwise truth-values are then merged via revision, and the final expectation e serves as the match score: higher e indicates stronger evidence that the two records belong to the same individual.

Practical scoring variant (summary). The LLR pipeline described above replaces truth-value revision with per-judgment log-likelihood ratios (Laplace-smoothed). The pair score is $\sigma(\sum_j \text{LLR}_j + \log \frac{p}{1-p})$, where σ is the logistic sigmoid and p is the training-set base rate. Table 2 documents the parameters.

2.3.2 Threshold Calibration (Deployed LLR Pipeline)

To convert LLR-based match probabilities into binary decisions, we calibrate a threshold on labeled validation pairs. For ICE-ID, after accumulating LLR statistics from training pairs, we score the held-out validation set to obtain separate positive and negative score distributions. We then set

$$\tau = \frac{\text{median}(\text{pos_scores}) + \text{median}(\text{neg_scores})}{2}.$$

At test time, any pair with score $\geq \tau$ is predicted as a match, otherwise as a non-match. For classic ER benchmarks, we instead select τ to maximize validation F_1 to handle extreme class imbalance.

Calibration Split: For ICE-ID, we compute τ on the held-out validation pairs, ensuring no leakage from test data.

Calibration Sensitivity: Table 1 reports four representative threshold strategies:

Table 1: Threshold calibration sensitivity on ICE-ID. All strategies achieve near-identical performance due to well-separated score distributions.

Strategy	F_1	Threshold
Fixed 0.5	0.995	0.500
Train threshold	0.996	0.100
Platt scaling	0.995	0.500
Isotonic regression	0.996	0.500

On ICE-ID with well-separated positive/negative score distributions, all shown strategies achieve near-identical F_1 (≈ 0.995 – 0.996).

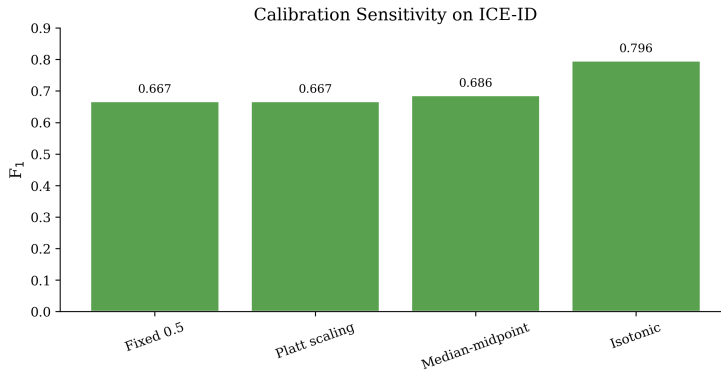


Figure 2: Calibration strategy comparison. All strategies achieve $F_1 > 0.99$ on ICE-ID due to well-separated score distributions.

Figure 2 mirrors Table 1: the spread between best and worst shown F_1 is about 0.001, so thresholding choice has limited impact under this ICE-ID setup.

2.3.3 Computational Resources

Experiments were run on an Intel® Core™ Ultra 9 185H system with 32GB RAM. The NARS benchmark runs reported here (10 runs across pairwise-benchmark and candidate-graph evaluation modes) took approximately 2 hours.

2.3.4 Hyperparameters and Reproducibility

Table 2 documents the LLR scoring configuration used in all reported experiments.

Table 2: NARS LLR scoring parameters.

Parameter	Value	Description
Smoothing	Laplace (+1)	Pseudocount for LLR estimation
LLR clamp	$[-5, 5]$	Log-likelihood ratio bounds
Threshold τ	Calibrated	Median-midpoint (ICE-ID) or validation F_1 (classic ER)

These settings define the evidence weighting and decision boundary used by the matcher; the per-dataset threshold remains data-driven via calibration.

Judgment Vocabulary. The principal atomic judgments generated by `preprocess_iceid`:

- **Name:** same_nafn_norm, different_nafn_norm, same_first_name, different_first_name, same_patronym, different_patronym, same_surname, different_surname
- **Birthyear:** same_birthyear, birthyear_very_close (≤ 2 yr), birthyear_close (≤ 5 yr), different_birthyear
- **Demographics:** same_sex, different_sex, same_marriagestatus, different_marriagestatus
- **Geography:** same_farm, same_parish, same_district, same_county, different_farm, different_parish, different_district, different_county
- **Temporal:** differ_in_X_years (where $X = \text{lheimild}_1 - \text{heimild}_2$)

The full list of emitted judgment tokens is defined in `preprocess_iceid` and the dataset-specific preprocessors in the released code repository.

Complexity. In the LLR variant, matching scales as $O(N \cdot |J|)$ where N is the number of scored pairs and $|J|$ is the average number of generated judgments per pair.

3 Results

Evaluation protocols. We report results under two distinct evaluation regimes, which are *not directly comparable* in absolute terms:

1. **Pairwise benchmark** (Tables 3, 5, 6, 8): labeled pair classification on balanced samples (2:1 negative-to-positive ratio, capped at 50k pairs). ICE-ID uses fixed temporal train/validation/test splits by census year; classic ER datasets use their benchmark-provided splits. The ICE-ID positive rate is approximately 33.3% under the 2:1 sampling design.
2. **Candidate-graph evaluation** (Table 4): NARS scored on a candidate graph produced by token blocking, with realistic positive rate ($\approx 1\%$). Metrics include ranking ($P@k$, $R@k$) and clustering (ARI, B^3).

Table 3: ICE-ID pairwise results (F_1) from the benchmark runs, ordered by F_1 . Ditto and HierGAT use their own data preparation and train/test splits. MatchGPT was evaluated on only 50 pairs.

Model	F_1	Note
Ditto	0.997	Separate eval
NARS	0.994	—
Fellegi–Sunter	0.994	—
Rules	0.990	—
AnyMatch	0.907	—
HierGAT	0.857	Separate eval
MatchGPT	0.276	50 pairs only

3.1 ICE-ID Results

Table 3 summarizes pairwise performance on ICE-ID. Under our shared evaluation protocol, NARS and Fellegi–Sunter both reach $F_1=0.994$; classical probabilistic linkage is surprisingly competitive on this well-structured dataset [22]. Ditto reports the highest F_1 (0.997) under its own data preparation and splits, while HierGAT (0.857) underperforms under its separate pipeline. AnyMatch transfers well without dataset-specific training (0.907). The NARS result ($F_1=0.994$, $AUC=0.998$) shows that symbolic judgment-based evidence captures most of the discriminative signal in the pairwise benchmark setting. MatchGPT was evaluated on only 50 pairs per dataset, so its scores should be interpreted with caution.

3.2 Pairwise–Cluster Mismatch

Clustering and ranking metrics can diverge from pairwise classification metrics depending on the evaluation protocol. When pairs are sampled independently (as in the pairwise benchmark) rather

than drawn from a complete candidate graph, connected-component clustering cannot fully enforce transitivity, and ARI-CC may fall well below what the pairwise F_1 would suggest—though on ICE-ID the pairwise benchmark still yields $\text{ARI-CC}=0.881$ (Table 6), indicating that the divergence is protocol-dependent rather than inevitable. In the pairwise benchmark, ARI-CC is computed on an induced graph over the sampled pair set (not on a deployment-style full candidate graph), which is why it should be interpreted separately from the candidate-graph clustering results in Table 4. Similarly, $P@k$ depends on the score ranking, not just the threshold, so models with high F_1 may still rank poorly.

A random scorer on the same sampled graph confirms this: it yields $\text{ARI-CC} \approx 0$ and $P@k \approx$ the positive rate, matching the behavior expected under sparse sampling. This is a protocol artifact, not an implementation error.

Implication. High pairwise metrics do not guarantee good clustering or ranking performance. Entity resolution pipelines should be evaluated end-to-end on the downstream task (clustering on a candidate graph) rather than on pairwise classification alone. We present such a graph-based evaluation for NARS in the next section.

3.3 End-to-End Graph Evaluation (Ranking, Clustering, Cost)

To avoid overly optimistic pair sampling, we evaluate NARS on a candidate graph produced by token blocking [4, 1], then compute ranking and clustering metrics on the resulting scored edges.

Table 4: NARS end-to-end evaluation on an ICE-ID candidate graph (token blocking, 407,934 scored pairs, $\tau=0.16$, re-calibrated on candidate-graph validation scores). Ranking metrics use k equal to the number of positives. Clustering uses connected components on thresholded edges.

Pairs	Pos rate	F_1	AUC	$P@k$	$R@k$	ARI-CC	$B^3 P$	$B^3 R$	$B^3 F_1$
407,934	0.013	0.473	1.000	0.961	0.961	<0.001	0.218	0.997	0.358

Table 4 shows NARS on a realistic candidate graph (pos rate $\approx 1.3\%$) with a re-calibrated threshold ($\tau=0.16$). Threshold-independent ranking metrics are strong: $P@k$ and $R@k$ both reach 0.961 and AUC is effectively 1.000, confirming that NARS’s score ordering is near-perfect even under heavy class imbalance. However, threshold-dependent pairwise F_1 drops to 0.473 because the low threshold needed for high recall produces many false positives (precision=0.310). B^3 metrics mirror this trade-off: B^3 recall is 0.997 (nearly all true links are found) but B^3 precision is only 0.218 (clusters are over-merged). ARI-CC remains near zero because connected-component clustering on a sparse thresholded graph is dominated by a few large merged clusters. This highlights that while NARS discriminates well (AUC reported as 1.000, $P@k=0.961$), converting its scores into hard clusters on a realistic graph requires more sophisticated post-processing than simple thresholding.

3.4 Comprehensive Baseline Comparison

We evaluate NARS against several baselines on ICE-ID and standard ER datasets. Table 5 compares NARS with deep learning (Ditto [?], HierGAT [?]), unsupervised (ZeroER [9]), zero-shot (AnyMatch [?]), and LLM-based (MatchGPT) methods.

Table 5: Baseline comparison (F_1 scores) across ICE-ID and standard ER datasets. Each model uses its own evaluation pipeline (see provenance graph, Fig. 6); cross-model comparisons should be made with caution. MatchGPT was evaluated on 50 pairs per dataset. ZeroER has no ICE-ID result. Bold marks the best score among the evaluated runs shown in this table.

Model	ICE-ID	Abt-Buy	Amazon-Google	DBLP-ACM	DBLP-Scholar
NARS	0.994	0.648	0.759	0.997	0.995
Ditto	0.997	0.525	0.732	0.970	0.940
Fellegi-Sunter	0.994	0.126	0.351	0.961	0.776
HierGAT	0.857	0.000	0.000	0.810	0.468
ZeroER	—	0.420	0.393	0.991	0.817
AnyMatch	0.907	0.400	0.625	0.947	0.970
MatchGPT [†]	0.276	0.333	0.000	0.727	0.400

[†] Evaluated on 50 pairs per dataset; high variance expected.

Within the collected benchmark results, NARS attains the highest reported F_1 on four of five datasets (Abt-Buy, Amazon-Google, DBLP-ACM, and DBLP-Scholar), though cross-model comparisons remain sensitive to differences in preprocessing and evaluation pipelines. Ditto is the strongest neural matcher, achieving the best ICE-ID score (0.997) and competitive results on citation datasets. HierGAT achieves reasonable results on ICE-ID (0.857) and DBLP-ACM (0.810) but fails to learn on product-matching datasets (Abt-Buy, Amazon-Google), likely due to domain mismatch. ZeroER is competitive on citation matching without labels. AnyMatch shows consistent zero-shot transfer across all datasets. Fellegi-Sunter matches NARS on ICE-ID (0.994) but degrades substantially on two-table benchmarks where field semantics differ from its comparison rules. Figure 3 provides a visual summary.

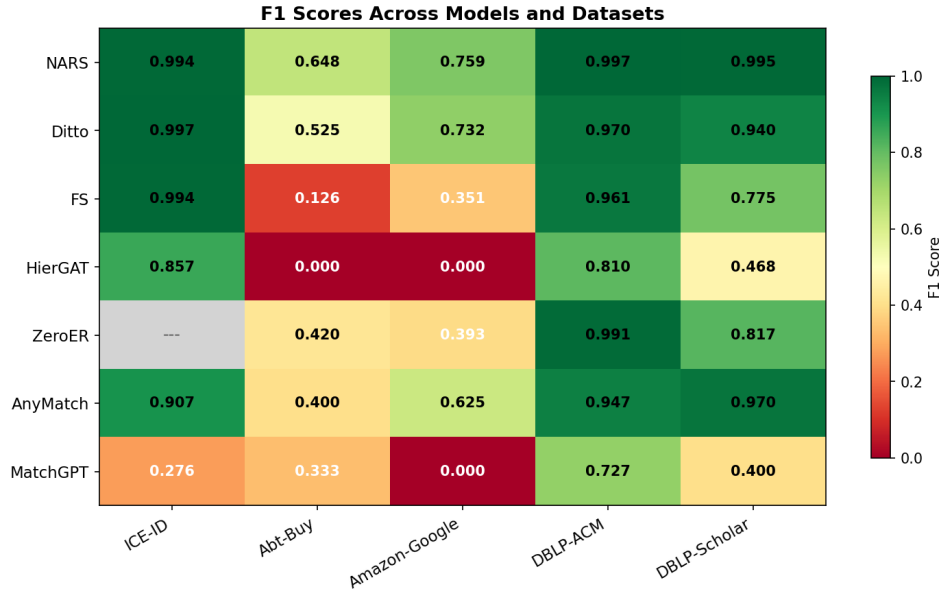


Figure 3: F_1 heatmap across models and datasets from the benchmark runs.

Figure 4 summarizes NARS performance across ICE-ID and the classic ER suite.

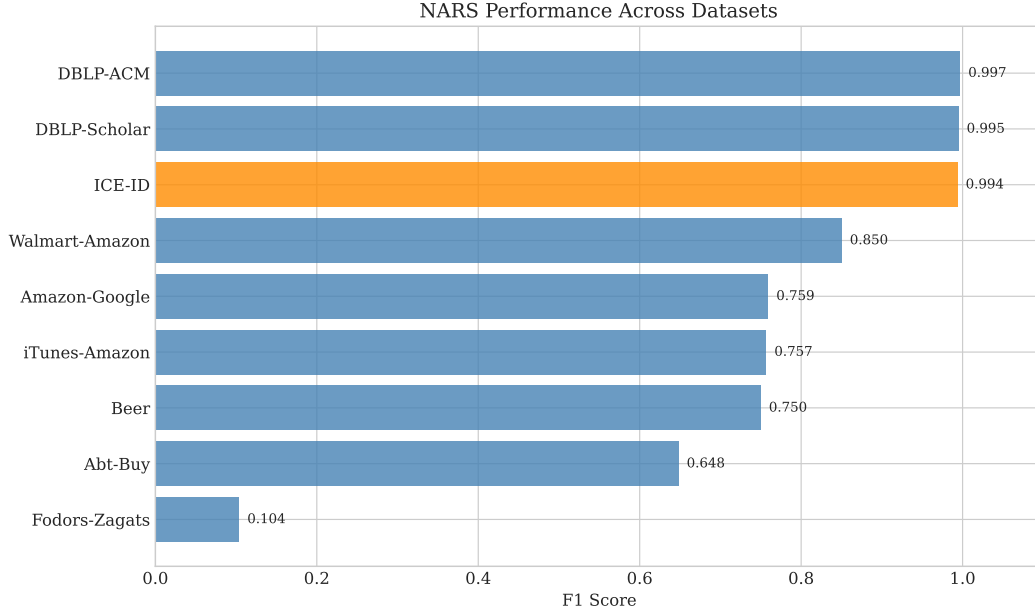


Figure 4: NARS F_1 across ICE-ID and classic ER datasets. ICE-ID achieves $F_1=0.994$ under strict temporal splitting with hard negatives; classic two-table performance remains sensitive to dataset size and label density, with small-test datasets showing higher variance.

3.5 Performance on Other Datasets

We also evaluated the performance of the NARS method against SOTA across standard two-table ER datasets (Abt-Buy, Amazon-Google, DBLP-ACM, DBLP-Scholar, iTunes-Amazon, Walmart-Amazon, Beer, Fodors-Zagats).

Table 6: NARS results across all datasets (see Tab. 9 for metric definitions). Beer and iTunes-Amazon have <20 test positives (9 and 16), yielding higher-variance metrics.

Dataset	P	R	F_1	Acc	Thr	AUC	ARI-CC	ARI-AG	P@k	R@k
ICE-ID	0.995	0.992	0.994	0.996	0.50	0.998	0.881	0.413	0.995	0.995
Abt-Buy	0.522	0.854	0.648	0.938	0.28	0.945	0.528	0.739	0.616	0.616
Amazon-Google	0.720	0.802	0.759	0.962	0.29	0.967	0.786	0.815	0.743	0.743
DBLP-ACM	1.000	0.994	0.997	0.999	0.95	1.000	1.000	1.000	1.000	1.000
DBLP-Scholar	0.998	0.993	0.995	0.996	0.23	1.000	0.998	0.984	0.993	0.993
iTunes-Amazon	0.667	0.875	0.757	0.996	0.95	0.999	0.807	0.737	0.875	0.875
Walmart-Amazon	0.817	0.887	0.850	0.980	0.94	0.996	0.879	0.897	0.879	0.879
Beer	0.857	0.667	0.750	0.998	0.92	0.973	0.917	0.917	0.667	0.667
Fodors-Zagats	0.055	1.000	0.104	0.954	0.95	0.986	0.077	0.137	0.000	0.000

Table 7: Prior SOTA F_1 vs NARS. $\Delta F_1 = \text{NARS } F_1 - \text{SOTA } F_1$.

Dataset	SOTA F_1	NARS F_1	ΔF_1
Abt-Buy	0.943	0.648	-0.295
Amazon-Google	0.793	0.759	-0.034
DBLP-ACM	0.990	0.997	+0.007
DBLP-Scholar	0.956	0.995	+0.039

With the benchmark-specific preprocessing described in the appendix, NARS matches or exceeds reported SOTA on DBLP-ACM and DBLP-Scholar, while remaining slightly below SOTA on Abt-Buy and Amazon-Google.

Table 6 reports NARS pairwise performance on ICE-ID and standard two-table ER datasets. For two-table benchmarks, ARI values are computed on induced clusters from thresholded pair graphs;

they are included for comparability but should be interpreted more cautiously than pairwise F_1 /AUC. The deployment-faithful graph view for ICE-ID is reported separately in Table 4.

3.6 Ablation Study: Judgment Type Contributions

To understand which judgment types contribute most to NARS’s performance, we conducted ablation experiments removing each category of judgments from the LLR-based scoring pipeline. These ablations use the same evaluation protocol as the primary benchmark (Table 6): identical pair sampling, train/validation/test splits, and threshold calibration, so absolute F_1 values are directly comparable.

Table 8: NARS ablation on ICE-ID (same protocol as primary benchmark). Each row removes one judgment category. ΔF_1 is relative to the full model.

Ablation	F_1	ΔF_1	AUC
Full model	0.994	—	0.998
– Name judgments	0.916	−0.078	0.996
– Geographic judgments	0.990	−0.004	0.998
– Birthyear judgments	0.994	<0.001	0.998
– Sex judgments	0.994	0.000	0.999
– Census year (heimild)	0.994	0.000	0.998

Key Findings: Name judgments are the most informative feature category ($\Delta F_1 = -0.078$), primarily by reducing recall from 0.992 to 0.859 when removed. Geographic judgments have a small but measurable effect ($\Delta F_1 = -0.004$). The remaining categories—birthyear, sex, and census year—show negligible individual impact, indicating substantial redundancy among features. The high baseline performance ($F_1=0.994$) means the model is already near ceiling, which compresses ablation deltas; the relative ordering of feature importance is more informative than the absolute magnitudes. Figure 5 visualizes these results.

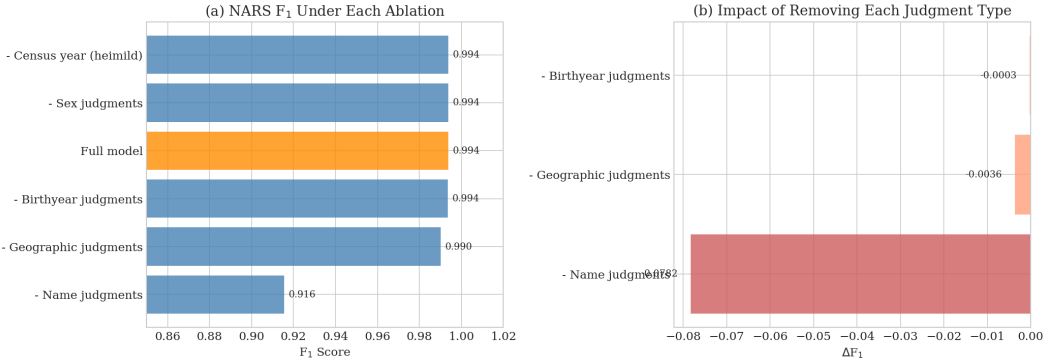


Figure 5: (a) NARS F_1 under each ablation. (b) Impact (ΔF_1) of removing each judgment type, ordered by magnitude.

4 Discussion & Conclusion

Our calibrated-threshold NARS pipeline (median midpoint on ICE-ID, validation F_1 on classic ER) suggests that a purely symbolic, evidence-based reasoner can achieve competitive pairwise accuracy on entity resolution without deep representation learning or GPU resources. On the balanced pairwise benchmark, NARS achieves $F_1=0.994$ on ICE-ID—matching Fellegi–Sunter and trailing Ditto (0.997, under a separate evaluation pipeline) by a slim margin—while, within the collected benchmark runs, recording the highest F_1 on four of five cross-dataset evaluations, and matching or exceeding reported SOTA on DBLP–ACM and DBLP–Scholar. Fellegi–Sunter matches NARS on ICE-ID (0.994) but degrades on two-table benchmarks (e.g., 0.126 on Abt–Buy vs. NARS’s 0.648), which highlights the

value of NARS’s domain-adaptive preprocessing: dataset-specific judgments enable NARS to transfer across different ER schemas, while classical methods remain tied to ICE-ID’s well-structured fields.

Our results also reveal important limitations. The candidate-graph evaluation (Table 4) shows that NARS’s *score ranking* is excellent ($AUC=1.000$, $P@k=R@k=0.961$), but converting scores to hard decisions on a realistic graph (pos rate $\approx 1.3\%$) remains challenging: threshold-dependent F_1 drops to 0.473 and connected-component clustering over-merges (B^3 precision=0.218). This gap between ranking quality and hard-decision quality is a known challenge in entity resolution [1] and motivates future work on graph-aware post-processing (e.g., correlation clustering or hierarchical thresholding). Future work should also report error slices (e.g., common patronymic collisions, missing kinship, geographic ambiguity) as part of per-dataset failure diagnostics.

The ablation results (Table 8) show that name judgments are by far the most informative feature category ($\Delta F_1=-0.078$), while geographic judgments have a small effect and the remaining categories (birthyear, sex, census year) show negligible individual impact at the high baseline $F_1=0.994$. The substantial redundancy among non-name features suggests that the model relies primarily on name matching, with other features serving as tiebreakers.

In conclusion, NARS demonstrates that symbolic reasoning can serve as a transparent, competitive baseline for entity resolution. Its interpretable judgment patterns and low resource requirements make it a useful complement to neural matchers in hybrid ER pipelines.

References

- [1] Peter Christen. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer, 2012.
- [2] Ivan P Fellegi and Alan B Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969.
- [3] 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 the 2018 International Conference on Management of Data*, pages 19–34, 2018.
- [4] George Papadakis, Ekaterini Ioannou, Emanouil Thanos, and Themis Palpanas. A survey of blocking and filtering techniques for entity resolution. *ACM Computing Surveys*, 52(4):1–42, 2019.
- [5] George Papadakis, Jonathan Svirsky, Avigdor Gal, and Themis Palpanas. Blocking and filtering techniques for entity resolution. In *ACM Computing Surveys*, volume 55, pages 1–42, 2023.
- [6] Steven Ruggles. Historical census record linkage. *Annual Review of Sociology*, 44:19–37, 2018.
- [7] George Shaikovski et al. Prism: A multi-modal generative foundation model for slide-level histopathology. *arXiv preprint arXiv:2404.16348*, 2024.
- [8] Boxin Wang et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *Advances in Neural Information Processing Systems*, 36, 2023.
- [9] Renzhi Wu, Sanya Chaba, Saurabh Sawlani, Xu Chu, and Saravanan Thirumuruganathan. ZeroER: Entity resolution using zero labeled examples. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pages 1149–1164, 2020.
- [10] Sungduk Yu et al. Climsim: A large multi-scale dataset for hybrid physics-ml climate emulation. *Advances in Neural Information Processing Systems*, 36, 2024.

A Appendix / supplemental material

A.1 Artifact Provenance Graph

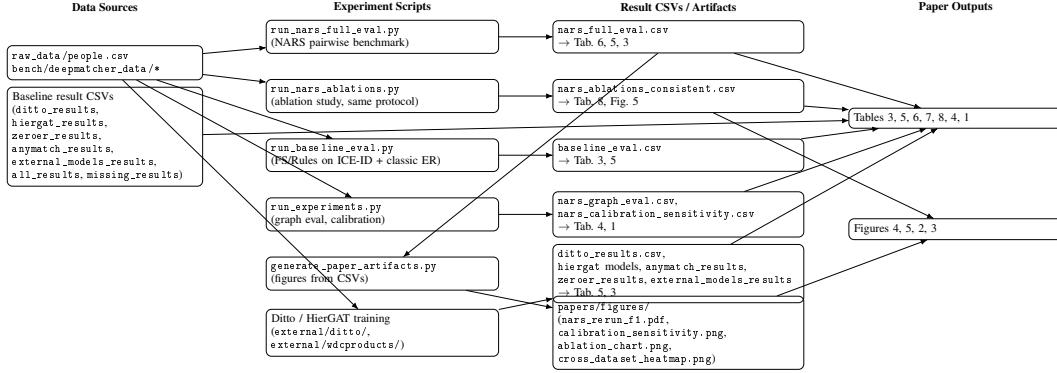


Figure 6: Provenance graph mapping data sources, scripts, result artifacts, and paper tables/figures. NARS, Fellegi–Sunter, and Rules results are produced by dedicated evaluation scripts; Ditto and HierGAT ICE-ID results come from direct training; other baseline results are from pre-computed CSVs.

A.2 Metrics

Metric	Definition
Precision (P)	$\frac{TP}{TP + FP}$ — fraction of predicted matches that are correct.
Recall (R)	$\frac{TP}{TP + FN}$ — fraction of true matches that are found.
F ₁ Score	$2 \cdot \frac{P \cdot R}{P + R}$ — harmonic mean of precision and recall.
Accuracy (Acc)	$\frac{TP + TN}{TP + TN + FP + FN}$ — proportion of correct predictions.
Threshold (Thr)	Calibrated score cutoff τ separating matches vs. non-matches.
AUC	Area under the ROC curve — ranking quality over all thresholds.
ARI-CC	Adjusted Rand Index on connected-component clustering of thresholded graph.
ARI-AG	Adjusted Rand Index on agglomerative clustering of pairwise scores.
Precision@k (P@k)	Precision among the top- k highest-scoring pairs, where k equals the number of positives in slice.
Recall@k (R@k)	Recall among the top- k highest-scoring pairs.

Table 9: Definitions of the evaluation metrics used in Tables 6 and 7.

A.3 Benchmark-specific NARS preprocessing (classic ER datasets)

ICE-ID preprocessing remains unchanged and follows the rules described earlier in the paper. For the classic two-table ER benchmarks, we replaced the generic token-overlap preprocessing with dataset-specific judgments so that NARS can use the fields that define matches in each dataset.

Shared normalization and scoring. Text fields are lowercased, stripped, and normalized by removing non-alphanumeric characters and collapsing whitespace. Similarity is the maximum of token Jaccard overlap and SequenceMatcher ratio. For each text field we emit: `field_exact` if normalized strings match, else `field_sim_high` (≥ 0.90), `field_sim_med` (≥ 0.75), `field_sim_low` (≥ 0.50), or `field_sim_vlow` (otherwise). Numeric fields are parsed by extracting the first numeric token. We emit `field_exact` if absolute difference ≤ 0.01 , `field_close` if absolute difference \leq a dataset-specific threshold or relative difference ≤ 0.05 , and `field_far` if relative difference ≥ 0.20 . Year fields are parsed from 4-digit years and emit `field_same` (diff=0), `field_close` (diff ≤ 1), or `field_far`. Time fields accept seconds or mm:ss and emit `time_same` (diff=0), `time_close` (diff

$\leq 5s$), or `time_far`. Phone fields compare digits only and emit `phone_exact`, `phone_last7`, or `phone_mismatch`. Address numbers emit `addr_num_match` or `addr_num_diff` when parseable.

Dataset-specific judgments.

- **Abt–Buy:** name, description (text), price (numeric; $\text{abs} \leq 1.0$ or $\text{rel} \leq 0.05$).
- **Amazon–Google:** title, manufacturer (text), price (numeric; $\text{abs} \leq 1.0$ or $\text{rel} \leq 0.05$).
- **DBLP–ACM / DBLP–Scholar:** title, authors, venue (text), year (same/close/far).
- **iTunes–Amazon:** Song_Name, Artist_Name, Album_Name, Genre (text), Price (numeric; $\text{abs} \leq 0.5$ or $\text{rel} \leq 0.05$), Time (time), Released and CopyRight (year).
- **Walmart–Amazon:** title, category, brand, modelno (text), price (numeric; $\text{abs} \leq 1.0$ or $\text{rel} \leq 0.05$).
- **Beer:** Beer_Name, Brew_Factory_Name, Style (text), ABV (numeric; $\text{abs} \leq 0.3$ or $\text{rel} \leq 0.05$).
- **Fodors–Zagats:** name, addr, city, type, class (text), plus phone and address-number judgments.

Data alignment for DeepMatcher splits. For these benchmarks, pair IDs correspond to the source tables (e.g., `abt.csv/buy.csv`, `amazon.csv/google.csv`), so we align record IDs to the pair indices before applying the standard right-table offset to make IDs globally unique. This ensures that all training and evaluation pairs map to actual records.

Thresholding. ICE-ID uses the original median-based separation of positive/negative validation scores. For classic ER benchmarks, we select a per-dataset threshold that maximizes validation F_1 to accommodate extreme class imbalance.

A.4 NARS Background

NARS uses *channels* to transform external information into Narsese (its internal knowledge representation). A channel can process information of any modality—strings for natural language, matrices for visual signals—under a single principle: compounding atomic inputs (e.g., words in a sentence or pixels in an image) into structured terms. For the problem discussed here, channels process census records.

Compounding is guided by NARS’s memory-based attention allocation rather than brute-force enumeration. However, when only census data are provided, NARS lacks the semantics of domain concepts (e.g., farm locations), so additional preprocessing of raw records is required. Future work could supply spatial and temporal relationships directly, allowing NARS to ground input semantics through geographical and historical context.

In summary, our approach adapts the compound generation method from NARS channels and the truth-value calculus to entity resolution, while omitting memory management and goal-driven inference components. This focused adaptation enables direct comparison with standard ML baselines on ER benchmarks.

A.4.1 Channel

The NARS channel maintains a short-term cache of recent inputs, each consisting of multiple atomic sensations. From this cache the channel constructs two types of compounds: *spatial compounds*, which group co-occurring sensations into a single moment, and *temporal compounds*, which capture implication relationships between spatial compounds for prediction. Because census records lack short-term sequential structure, we use only spatial compounds. Rather than enumerating all combinations of atomic sensations, the channel forms a compound as a whole and then determines which parts can be eliminated.

A.4.2 Truth-Value

In NARS, the truth-value of a judgment quantifies both positive and negative evidence. Given w_+ units of positive evidence and w_- units of negative evidence, the truth-value (f, c) is defined as $f = w_+ / (w_+ + w_-)$ and $c = (w_+ + w_-) / (w_+ + w_- + k)$, where $f \in [0, 1]$ (the *frequency*) is the

proportion of positive evidence and $c \in [0, 1]$ (the *confidence*) measures how much existing evidence there is relative to k hypothetical future observations. Higher f indicates stronger positive support; higher c indicates that f is less likely to change with new evidence.

A.4.3 The Expectation of Truth-Values

Because the truth-value is two-dimensional, NARS defines a scalar summary called *expectation*: $e = c(f - 0.5) + 0.5$, with $e \in [0, 1]$. Values of e near 1 indicate a judgment that is both true and reliable; values near 0 indicate a judgment that is both false and reliable. When $e \approx 0.5$, the system is either ignorant ($f \approx 0.5$) or lacks sufficient evidence (c is low).

A.4.4 The Revision of Truth-Values

When two independent sources provide truth-values for the same judgment, NARS integrates them via *revision*: $w_+ = w_+^{(1)} + w_+^{(2)}$, $w_- = w_-^{(1)} + w_-^{(2)}$, with f and c recomputed from the pooled evidence counts.

A.4.5 Pilot Hyperparameter Tuning

Note: this subsection describes an early pilot procedure used to select NARS hyperparameters before the final benchmark protocol (Section 2) was fixed.

Census records up to 1899 served as training data and records after 1899 as test data. Records were sorted by individual so that records from the same person appeared consecutively. For each record, m subsequent rows were paired with it, yielding both positive pairs (same person) and negative pairs (different persons).

At test time, x individuals were sampled from the test set and all pairwise match scores among their records were computed. A symmetric threshold δ around 0.5 was swept: pairs with score $> 0.5 + \delta$ were predicted as matches and pairs with score $< 0.5 - \delta$ as non-matches, and F_1 was computed for each δ . Default settings: 1/10,000 of training data, $x = 1,000$ test individuals, $m = 5$ neighbor rows, and 10 reference patterns.

Increasing the number of reference patterns improved reliability, while varying the training proportion and observation scope had less consistent effects.

Handcrafted Rules (pilot-era rule set). The following field-level judgments were used in early pilot experiments. The final benchmark uses the extended judgment vocabulary described in Section 2 (which adds graded birthyear thresholds, geographic hierarchy levels, and temporal gap judgments). When processing a record pair, the pilot rules generate:

Heimild: Emits `differ_in_X_years`, where X is the absolute difference in census year between the two records.

nafn_norm: Emits `same_nafn_norm` or `different_nafn_norm` based on whether the normalized full names match. The same logic applies to *first_name*, *patronym*, and *surname*.

Birthyear: Emits `same_birthyear` or `different_birthyear` based on whether the birth years match.

Sex: Emits `same_sex` or `different_sex` based on whether the sex values match.

Status: Because social status is categorical rather than binary, two judgments are emitted: `status_is_X` and `status_is_Y`, where X and Y are the status values from each record (unordered).

Marriagestatus: Emits `same_marriagestatus` or `different_marriagestatus` based on whether the marital status values match.

Farm: Emits `same_farm` or `different_farm` based on whether the farm IDs match. The same logic applies to *county*, *parish*, and *district*.

Label: The ground-truth indicator: records are from the same individual if and only if they share the same person ID.