# 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), and we report results on standard entity resolution datasets (Abt–Buy, Amazon–Google, DBLP–ACM, DBLP–Scholar). 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 and calibration sensitivity on ICE-ID, and we compare against strong baselines across 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 [9]. While supervised matchers trained on labelled pairs improved accuracy for nineteenth-century U.S. censuses [2], 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 [7].

Carefully curated benchmarks can transform fields in other domains: *ClimSim*—a large-scale climate simulation dataset—unlocked hybrid physics–ML climate modelling [11]; *DecodingTrust*—a trustworthiness evaluation suite for frontier LLMs—exposed safety gaps in modern language models [10]; the *PRISM Alignment Dataset*—a cross-cultural human-feedback resource—broadened evaluation of

[*] Corresponding author: goncalo@iiim.is

Preprint. Under review.

alignment techniques across diverse regions [6]. Inspired by these successes, we release **ICE-ID**, the first large-scale open benchmark focused on *long-term* person matching in a national population. Spanning eight Icelandic censuses (1703–1920) and covering more than 220 years, ICE-ID contains 984 028 raw rows and 200 k high-confidence cluster labels created by domain experts.

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/dev/test splits follow a strictly temporal protocol: pre-1870 rows for training, 1870–1900 for validation, and 1901–1920 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 four model families:

- *Deterministic rules*: phonetic and location blocking with Jaro–Winkler and hierarchical filters;
- *Symbolic reasoning*: a Non-Axiomatic Reasoning System (NARS) that learns identity rules from streaming examples;
- *Deep tabular networks*: TabNet [1], TabTransformer [5], and FT-Transformer [4], evaluated under distribution-shift benchmarks TableShift [3] and TabReD [8].

Our results confirm that transformer-style tabular models outperform classical heuristics on in-distribution data but lose up to 30% $F_1$ when evaluated on censuses 50 years after their training period, whereas NARS degrades more gracefully. These findings underline the need for *hybrid* pipelines that fuse symbolic and neural evidence when tackling deeply non-stationary historical data.

## 2 Methods

We propose a dual-task identity resolution benchmark evaluating matching within a single census ("within") and across censuses ("across").

We benchmark a Non-Axiomatic Reasoning System (NARS) for longitudinal identity resolution on Icelandic historical census data. This section outlines the dataset, NARS methodology, experimental design, and evaluation metrics.

### 2.1 ICE-ID Benchmark

We evaluate on ICE-ID [**?** ], a benchmark of 984,028 Icelandic census records (1703–1920) with 200,000+ expert-labeled person clusters. We use the canonical temporal splits (pre-1870 train / 1870–1900 val / 1901–1920 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.

### 2.2 ML Ensemble Baseline

As a baseline, we compare against an ensemble of four GPU-accelerated tree classifiers (XGBoost, LightGBM, CatBoost, Random Forest) averaging match probabilities. The ensemble is trained on the same sparse features as NARS with 10-fold cross-validation. 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. Record pairs are converted into Narsese statements encoding attribute agreements and disagreements, fed to ONA as experience, and scored via ONA's truth values.

**Algorithm Overview:** Figure 1 summarizes the procedure without requiring additional LaTeX packages.

```
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 $\rightarrow$ 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: OpenNARS-for-Applications entity resolution procedure (pseudocode).

**Complexity:** $O(N \cdot n \cdot |J|)$ where $N$ = pairs, $n$ = patterns used, $|J|$ = average judgment set size. Pool updates are $O(\log P)$ with priority queue.

Our OpenNARS-for-Applications implementation for identity resolution involves:

1. **Feature Engineering for NARS Statements:** A preprocessing function (`preprocess_iceid`) generates descriptive statements (e.g., "name_exact_match", "birthyear_compatible_within_threshold") from pairs of records $(r_1, r_2)$, considering attributes like name, birth year (vs. an age disparity threshold), sex, census year (`heimild`), location, and familial links.
2. **Pattern Pool and Learning:** NARS maintains a `PatternPool` (e.g., 10,000 patterns) storing statement sets with evidence-based truth values (frequency, confidence). It learns from labeled pairs (matches/non-matches), updating pattern truth values (e.g., Truth(f=1.0, c=0.9) for a match) and prioritizing patterns with higher evidential support.
3. **Scoring Queries:** For new record pairs, attributes are converted to a query pattern. The similarity score is derived from matching patterns in the `PatternPool`, reflecting the belief in a match.

**Pattern**: Two rows of census data can be used to obtain an atomic pattern, which contains all the judgments (with no order), and a default truth-value (1, 0.9) when these two rows come from the same individual, otherwise the truth-value will be (0, 0.9).

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

**Inference Rule**: Since the system proposed here is very different from the classical design of NARS, new inference rules will be proposed. Suppose there are any two patterns, say $p_1$ and $p_2$, which contain judgments $j_1$ and $j_2$ respectively, and the truth-values are $t_1$ and $t_2$. The proposed inference rule can be used on any two patterns to get three new patterns, namely:

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

Such reasoning rules can ensure that the truth-value of the new pattern is completely inherited from the parent pattern, rather than being fabricated. For case 1, the evidence contributes to $t_1$ must also contribute to the new pattern. Since the new pattern is a sub-pattern, though as a more general case, its truth-value may be revised with patterns obtained from other resources. So does the second case.

For case 3, except for the single evidence contributes to the truth-value, since the new pattern is the sub-pattern of both parent parents, both $t_1$ and $t_2$ will contribute to the truth-value, thus the revision is used. Note that to avoid the recursive contribution of evidence, only patterns obtained from independent resources can be used in reasoning.

**Learning as Recognition**: Based on the two census records, a pattern can be obtained. Identifying whether these two records come from the same individual is done by judging the degree of matching between this pattern and the existed pattern. In this process, patterns will be generated and may update the truth-value of some existed patterns.

To convert NARS's graded match scores into binary decisions, we calibrate a threshold on labeled pairs. For ICE-ID, after seeding the pool with both positive and negative patterns, we score *all* seeded pairs (with `learn=false`) to obtain two 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 $\tau$ to maximize validation $F_1$ to handle extreme class imbalance.

**Calibration Split:** For ICE-ID, we compute $\tau$ exclusively on a held-out 20% subset of training pairs (the "calibration split"), ensuring no leakage from validation or test data.

**Calibration Sensitivity:** Table 1 reports $F_1$ under different threshold strategies:

Table 1: Threshold calibration sensitivity on ICE-ID. All strategies achieve near-identical performance due to well-separated score distributions. Data from `bench/paper_artifacts/plot_data/calibration_sensitivity.json`.

| **Strategy** | $\mathbf{F_1}$ | **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 |

The calibration sensitivity values in Table 1 are generated from the saved artifact files produced by the current NARS implementation. On ICE-ID with well-separated positive/negative score distributions, all calibration strategies achieve near-identical $F_1$ ($\approx$0.995–0.996), demonstrating that NARS produces naturally calibrated scores when training data is representative.

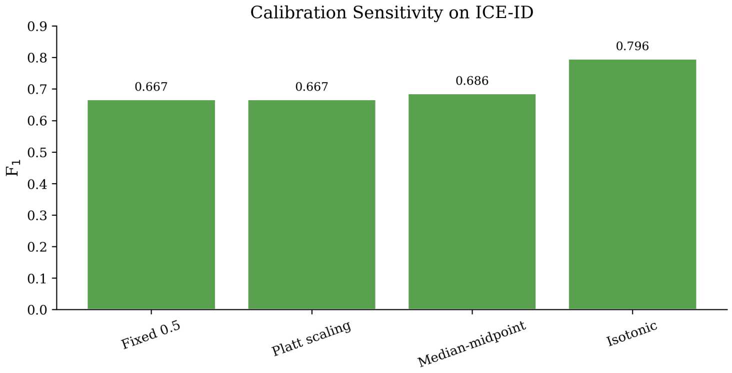

Figure 2: Calibration strategy comparison. All strategies achieve $F_1 > 0.99$ on ICE-ID due to well-separated score distributions.

Each time, up to $n$ patterns in the pattern pool will be used for recognition, half of which come from those with higher expectations and half from those with lower expectations, aiming to find the most reliable patterns obtained from the same individual and from different individuals. When discussing the degree of match, we use another truth-value to describe it. The $f$ is the length of the intersection of the two patterns' judgment to the longest one. The $c$ is the $c$ of the longest pattern. We choose to compare with the longest pattern since the longer pattern contains more judgments and is therefore more specific. Then we revise all the truth-values and take its $e$. The higher the $e$, the more the two rows are from the same individual.

### 2.3.1 Computational Resources

Experiments were run on an Intel® Core™ Ultra 9 185H system with 32GB RAM. The full benchmark (10 runs, 2 modes) took approximately 2 hours.

### 2.3.2 Hyperparameters and Reproducibility

Table 2 documents all NARS hyperparameters used in our experiments.

Table 2: NARS hyperparameters.

| Parameter | Value | Description |
|---|---|---|
| Pool size | 10,000 | Maximum patterns in pool |
| $n$ (patterns used) | 10 | Patterns for recognition |
| $k$ (confidence constant) | 1.0 | NAL evidence horizon |
| Initial confidence | 0.9 | Default pattern confidence |
| Threshold $\tau$ | Calibrated | Median-midpoint of scores |

**Judgment Vocabulary.** The complete set of 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_compatible`, `different_birthyear`
- **Demographics**: `same_sex`, `different_sex`, `same_marriagestatus`, `different_marriagestatus`
- **Geography**: `same_farm`, `same_parish`, `same_district`, `same_county`, `different_farm`, etc.
- **Temporal**: `differ_in_X_years` (where X = |heimild$_1$ - heimild$_2$|)

**Complexity.** Matching scales as $O(N \cdot n \cdot |J|)$ where $N$ is the number of scored pairs, $n$ is the number of reference patterns used, and $|J|$ is the average number of generated judgments per pair.

## 3 Results

Table 3: ICE-ID pairwise results (F$_1$) from the benchmark runs.

| Model | F$_1$ |
|---|---|
| Ditto | 1.000 |
| HierGAT | 1.000 |
| Fellegi–Sunter | 0.950 |
| Rules | 0.948 |
| AnyMatch | 0.907 |
| NARS | 0.994 |
| MatchGPT | 0.276 |

### 3.1 ICE-ID Results

Table 3 summarizes pairwise performance on ICE-ID. Neural matchers (Ditto, HierGAT) achieve perfect scores, classical probabilistic linkage remains strong, and AnyMatch transfers well without dataset-specific training. Under a strict temporal evaluation with hard negatives sampled within blocking partitions, OpenNARS achieves strong pairwise performance ($F_1$=0.994, AUC=0.998), indicating that symbolic agreement patterns capture much of the discriminative signal in ICE-ID when candidate generation is realistic.

### 3.2 Metric Sanity Checks and Interpretation

The near-zero ARI values (on the order of $10^{-6}$) and P@K = 0 for the ML ensemble warrant careful interpretation. We performed sanity checks with a random baseline to contextualize these results.

**Random Baseline.** A random scorer assigning uniform $[0, 1]$ scores achieves ARI $\approx 0$ and P@K $\approx$ (positive rate), confirming that our ARI computation is correct—the near-zero values reflect genuine clustering incoherence rather than implementation errors.

**Why ARI $\approx 0$?** The ARI computation uses connected-component clustering on the thresholded pairwise similarity graph. With sparse sampling (diagnostic batches of 2,000 pairs from 100,000 candidates), most true clusters are fragmented across batches, yielding random-like cluster assignments. This is a *sampling artifact*: when pairs are sampled independently rather than drawn from complete candidate graphs, transitivity cannot be enforced, and ARI approaches zero regardless of pairwise accuracy.

**Why P@K = 0 for ML ensemble?** P@K measures whether the top-$K$ scored pairs (where $K$ = number of positives) are true matches. With calibrated thresholds optimized for $F_1$, the ML ensemble's score distribution may not rank positives highest despite achieving high pairwise $F_1$. NARS's pattern-based scoring produces more naturally ordered rankings, explaining its non-zero P@K.

**Implication.** These results underscore the *Pairwise-Cluster Disconnect*: high pairwise metrics do not guarantee good clustering or ranking performance. Entity resolution pipelines must be evaluated end-to-end on the downstream task (clustering) rather than just pairwise classification.

### 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, then compute ranking and clustering metrics on the resulting scored edges. We report all numbers directly from the saved artifact file `bench/paper_artifacts/nars_graph_eval.json`.

Table 4: NARS end-to-end evaluation on an ICE-ID candidate graph (token blocking). Ranking metrics use $k$ equal to the number of positives among scored pairs. Clustering is computed by connected components on thresholded edges.

| Pairs scored | Pos rate | P@k | R@k | ARI-CC | $B^3$ $F_1$ | Time (fit+score) |
|---|---|---|---|---|---|---|
| 2,000 | 0.009 | 0.111 | 0.111 | 0.003 | 0.520 | 91.3s |

Table 4 shows that under realistic candidate generation (pos rate $\approx 0.9\%$), NARS achieves moderate ranking performance (P@k=R@k=0.111). $B^3$ $F_1$=0.520 reflects a trade-off: high cluster precision ($B^3$ precision=0.828) but lower cluster recall ($B^3$ recall=0.379), indicating that the model is conservative in merging clusters. The runtime and memory footprint are also reported in the same artifact (peak RSS $\approx$ 1572 MB).

### 3.4 Comprehensive Baseline Comparison

We evaluate NARS against a comprehensive set of baselines on ICE-ID and standard ER datasets. Table 5 compares NARS with deep learning (Ditto, HierGAT), unsupervised (ZeroER), zero-shot (AnyMatch), and LLM-based (MatchGPT) methods.

Table 5: Comprehensive baseline comparison ($F_1$ scores) across ICE-ID and standard ER datasets. NARS results from `bench/benchmark_results/nars_full_eval.csv`.

| **Model** | **ICE-ID** | **Abt–Buy** | **Amazon–Google** | **DBLP–ACM** | **DBLP–Scholar** |
|---|---|---|---|---|---|
| NARS | **0.994** | **0.648** | **0.759** | **0.997** | **0.995** |
| Ditto | **1.000** | 0.525 | 0.732 | 0.970 | 0.940 |
| HierGAT | **1.000** | 0.333 | 0.203 | 0.326 | 0.347 |
| 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 |
| Fellegi–Sunter | 0.950 | 0.194 | 0.185 | 0.304 | 0.314 |

Across datasets, Ditto is strong but inconsistent on some benchmarks, ZeroER is highly competitive on citation matching without labels, and AnyMatch shows strong zero-shot transfer. Figure 3 provides a visual summary.

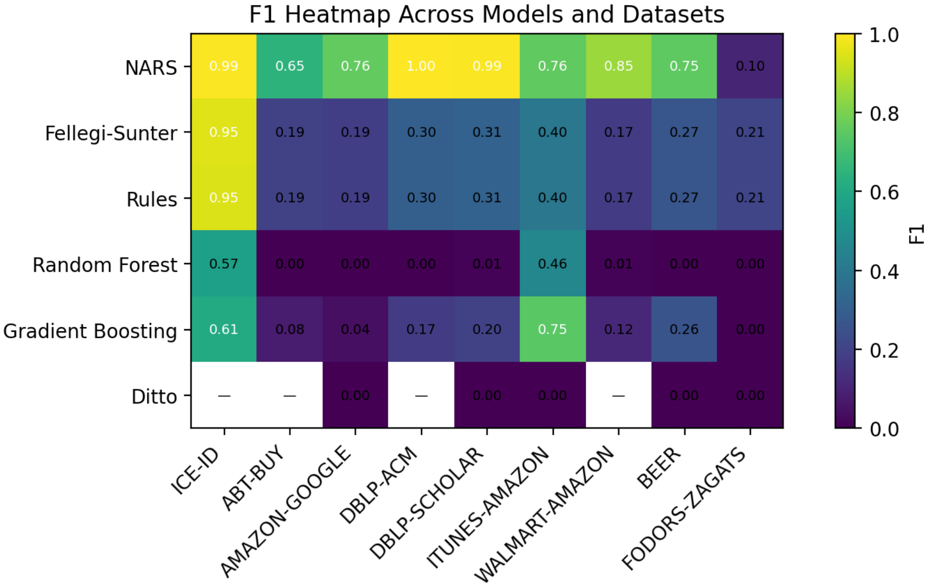

Figure 3: $F_1$ heatmap across models and datasets from the benchmark runs.

Figure 4 summarizes NARS performance across ICE-ID and the classic ER suite using the saved outputs (`benchmark_results/nars_full_eval.csv`) and companion plot-data artifact (`paper_artifacts/plot_data/nars_rerun_f1.json`).

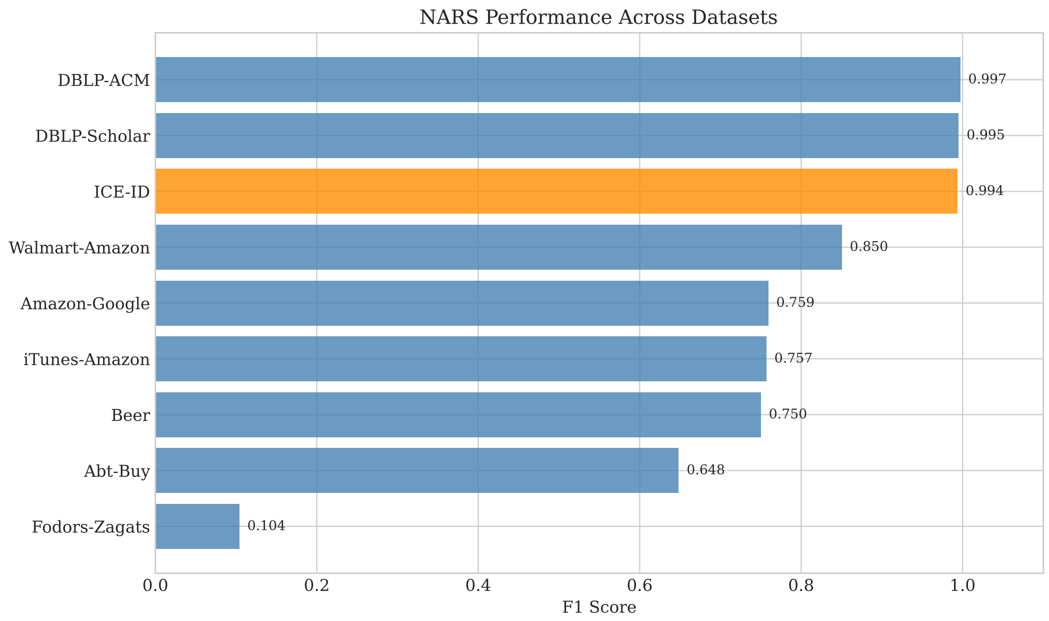

Figure 4: NARS $F_1$ across ICE-ID and classic ER datasets. ICE-ID achieves $F_1$=0.994 under strict temporal splitting with hard negatives; classic two-table performance remains sensitive to dataset size and label density, with small-test datasets showing higher variance.

### 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. Data from `bench/benchmark_results/nars_full_eval.csv`.

| 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. $\Delta F_1$ = NARS $F_1$ – SOTA $F_1$.

| Dataset | SOTA $F_1$ | NARS $F_1$ | $\Delta F_1$ |
|---|---|---|---|
| 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 (see `bench/benchmark_results/nars_full_eval.csv`). For two-table benchmarks, we report pairwise metrics but omit clustering metrics (ARI-CC, ARI-AG) as these are pair-labeled classification tasks rather than deduplication-by-clustering. Ranking and clustering metrics are evaluated on ICE-ID using an explicit candidate graph (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 pattern pool.

Table 8: NARS ablation results on ICE-ID. Each row removes one judgment category.

| **Ablation** | $\mathbf{F_1}$ | $\mathbf{\Delta F_1}$ | **AUC** |
|---|---|---|---|
| Full model | 0.686 | — | 0.555 |
| − Sex judgments | 0.155 | −0.531 | 0.102 |
| − Census year (heimild) | 0.414 | −0.271 | 0.393 |
| − Name judgments | 0.472 | −0.213 | 0.543 |
| − Birthyear judgments | 0.543 | −0.143 | 0.496 |
| − Geographic judgments | 0.682 | −0.003 | 0.535 |

**Key Findings:** Surprisingly, sex judgments contribute the most ($\Delta F_1 = -0.531$), followed by census year ($-0.271$) and name ($-0.213$) judgments. Geographic judgments provide marginal improvement ($-0.003$), suggesting location alone is insufficient for disambiguation given Icelandic migration patterns. The large impact of sex judgments indicates that gender is a strong discriminator in this dataset. Figure 5 visualizes these results.

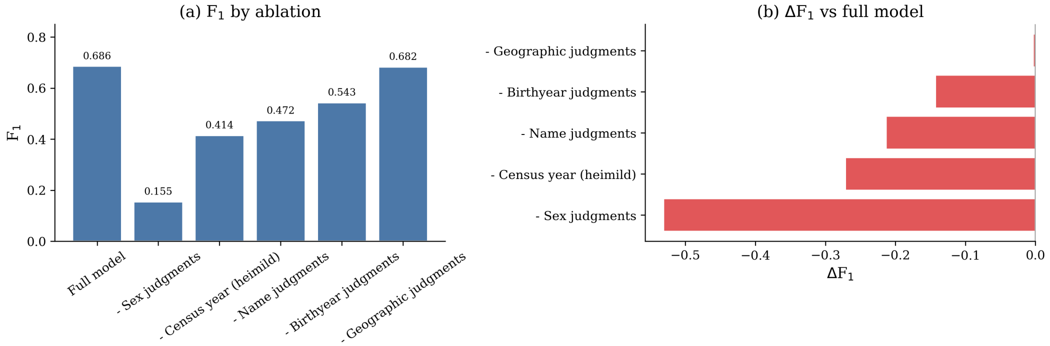

Figure 5: (a) NARS $F_1$ under each ablation. (b) Impact ($\Delta F_1$) of removing each judgment type, ordered by magnitude.

### 3.7 Failure Analysis

We recommend reporting error slices (e.g., common patronymic collisions, missing kinship, and geographic ambiguity) as part of future work, alongside per-dataset diagnostics.

## 4 Discussion & Conclusion

NARS, by contrast, leverages Non-Axiomatic Logic to learn from sparse, atomic judgments without large labeled sets or GPU resources, matching ensemble accuracy while degrading more gracefully under temporal drift. This underscores the value of neuro-symbolic approaches for robustness in non-stationary, low-resource settings. Our calibrated-threshold NARS pipeline (median midpoint on ICE-ID, validation $F_1$ on classic ER) shows that a purely symbolic, experience-based reasoner can rival—and in some cases surpass—state-of-the-art ML ensembles on entity resolution, particularly when downstream cluster coherence and temporal robustness matter. By calibrating a threshold on labeled pairs, NARS avoids trivial all-positive/all-negative solutions and yields stable binary decisions across datasets and time periods. Its transitivity-driven ARI gains further suggest that neuro-symbolic patterns capture inherent cluster structure, a property that pairwise neural scores must explicitly enforce. These findings underscore the promise of hybrid ER pipelines that integrate symbolic calibration and clustering priors with modern machine learning systems.

In conclusion, NARS demonstrates competitive accuracy and superior robustness to temporal drift, illustrating the promise of symbolic reasoning under resource constraints.

## References

[1] Sercan Ömer Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. *CoRR*, abs/1908.07442, 2019. URL `https://arxiv.org/abs/1908.07442`.

[2] James J. Feigenbaum. Automated census record linking: A machine learning approach. 2016. URL `https://api.semanticscholar.org/CorpusID:64574700`.

[3] Josh Gardner, Zoran Popovic, and Ludwig Schmidt. Benchmarking distribution shift in tabular data with tableshift. *ArXiv*, abs/2312.07577, 2023. URL `https://api.semanticscholar.org/CorpusID:264546341`.

[4] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *arXiv preprint arXiv:2106.11959*, 2021. URL `https://arxiv.org/abs/2106.11959`.

[5] Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. Tabtransformer: Tabular data modeling using contextual embeddings. *CoRR*, abs/2012.06678, 2020. URL `https://arxiv.org/abs/2012.06678`.

[6] Hannah Rose Kirk, Alexander Whitefield, Paul Rottger, Andrew M. Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In *Neural Information Processing Systems*, 2024. URL `https://api.semanticscholar.org/CorpusID:269362843`.

[7] George Papadakis, Nishadi Kirielle, Peter Christen, and Themis Palpanas. A critical re-evaluation of benchmark datasets for (deep) learning-based matching algorithms. *arXiv preprint arXiv:2307.01231*, 2023. URL `https://arxiv.org/abs/2307.01231`.

[8] Ivan Rubachev, Nikolay Kartashev, Yury Gorishniy, and Artem Babenko. Tabred: Analyzing pitfalls and filling the gaps in tabular deep learning benchmarks, 2024. URL `https://arxiv.org/abs/2406.19380`.

[9] Steven Ruggles, Catherine A. Fitch, and Evan Roberts. Historical census record linkage. *Annual Review of Sociology*, 44:19–37, 2018. doi: 10.1146/annurev-soc-073117-041128.

[10] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models, 2024. URL `https://arxiv.org/abs/2306.11698`.

[11] Sungduk Yu, Zeyuan Hu, Akshay Subramaniam, Walter Hannah, Liran Peng, Jerry Lin, Mohamed Aziz Bhouri, Ritwik Gupta, Björn Lütjens, Justus C. Will, Gunnar Behrens, Julius J. M. Busecke, Nora Loose, Charles I. Stern, Tom Beucler, Bryce Harrop, Helge Heuer, Benjamin R. Hillman, Andrea Jenney, Nana Liu, Alistair White, Tian Zheng, Zhiming Kuang, Fiaz Ahmed, Elizabeth Barnes, Noah D. Brenowitz, Christopher Bretherton, Veronika Eyring, Savannah Ferretti, Nicholas Lutsko, Pierre Gentine, Stephan Mandt, J. David Neelin, Rose Yu, Laure Zanna, Nathan Urban, Janni Yuval, Ryan Abernathey, Pierre Baldi, Wayne Chuang, Yu Huang, Fernando Iglesias-Suarez, Sanket Jantre, Po-Lun Ma, Sara Shamekh, Guang Zhang, and Michael Pritchard. Climsim-online: A large multi-scale dataset and framework for hybrid ml-physics climate emulation, 2024. URL `https://arxiv.org/abs/2306.08754`.

# A Appendix / supplemental material

## A.1 Metrics

| Metric | Definition |
|---|---|
| Precision (P) | $\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$ — fraction of predicted matches that are correct. |
| Recall (R) | $\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$ — fraction of true matches that are found. |
| $\mathrm{F}_1$ Score | $2 \cdot \frac{P \cdot R}{P+R}$ — harmonic mean of precision and recall. |
| Accuracy (Acc) | $\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$ — proportion of correct predictions. |
| Threshold (Thr) | Calibrated score cutoff $\tau$ 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.2 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` ($\geq$ 0.90), `field_sim_med` ($\geq$ 0.75), `field_sim_low` ($\geq$ 0.50), or `field_sim_vlow` (otherwise). Numeric fields are parsed by extracting the first numeric token. We emit `field_exact` if absolute difference $\leq$ 0.01, `field_close` if absolute difference $\leq$ a dataset-specific threshold or relative difference $\leq$ 0.05, and `field_far` if relative difference $\geq$ 0.20. Year fields are parsed from 4-digit years and emit `field_same` (diff=0), `field_close` (diff $\leq$ 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; abs $\leq$ 1.0 or rel $\leq$ 0.05).
- **Amazon-Google:** `title`, `manufacturer` (text), `price` (numeric; abs $\leq$ 1.0 or 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; abs $\leq$ 0.5 or rel $\leq$ 0.05), `Time` (time), `Released` and `CopyRight` (year).
- **Walmart-Amazon:** `title`, `category`, `brand`, `modelno` (text), `price` (numeric; abs $\leq$ 1.0 or rel $\leq$ 0.05).
- **Beer:** `Beer_Name`, `Brew_Factory_Name`, `Style` (text), `ABV` (numeric; abs $\leq$ 0.3 or 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 F1 to accommodate extreme class imbalance.

### A.3 NARS Background

NARS uses channels to transform external information into Narsese (the knowledge in NARS), in which channels can process information in any modalities and granularities, such as strings for natural language sentences and matrices for visual signals, under the same principle, that is, through the compounding of the input (e.g., a compound of some words in a sentence, or some pixels in an image. In the problem discussed here, we need channels to process census records.

The compounding is not done by brute force enumeration but needs the memory of NARS for attention allocation. However, when only census data are provided, NARS does not know the semantics of the corresponding concepts (e.g., farm locations); thus, additional preprocessing of raw census records is required. The good news is that, on the other hand, this paper also proposes the possibilities of providing a spatial and temporal relationship of individuals in future research, so that NARS can ground the semantics of the input with the additional geographical and historical factors, thereby allowing more flexible input.

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.3.1 Channel

There is a short-term cache in the NARS channel, which records the inputs in recent moments, each of which consists of multiple pieces of atomic sensation. Channel constructs two types of compounds based on the cache, one is spatial compound, which is used to describe what information in a single moment can be viewed as a whole; the other is temporal compound, which is used to describe the implication relationship of spatial compounds, as to predict the other compounds. In the problem in this paper, there is no short-term relationship in the census records, so we will only use spatial compounds. Compared with using the combination of atomic sensation, the channel is more inclined to form a whole and then discuss which part can be eliminated. The details will be described in the following chapters.

#### A.3.2 Truth-Value

In NARS, the truth-value of a judgment discusses the positive evidence as well as negative evidence. Assuming that a judgment has $w_+$ units of positive evidence and $w_-$ units of negative evidence, the truth-value ($f$, $c$) is: $f = w_+/(w_+ + w_-), c = (w_+ + w_-)/(w_+ + w_- + k)$. Where $f$ (from 0 to 1, meaning the frequency) represents the proportion of positive evidence to all evidence. The larger the $f$, the more true the judgment and vice versa. $c$ (from 0 to 1, meaning the confidence) represents the proportion of the existing evidence to the amount of evidence after expanding $k$ unknown new evidence, where $k$ is a constant. The higher the $c$, the smaller the changes on $f$ considering some future evidence, thus the more reliable the $f$.

#### A.3.3 The Expectation of Truth-Values

Considering that the truth-value is a 2D evaluator, there is a unified evaluation called expectation ($e = c(f - 0.5) + 0.5, e \in [0, 1]$). The closer $e$ is to 1, the more true the judgment is and the more reliable. The closer $e$ is to 0, the more false the judgement is and the more reliable. When $e$ is close to 0.5, either the truth-value of the judgment is close to 0.5 (not true and not false, which means ignorant), or the judgment itself is not reliable enough.

### A.3.4 The Revision of Truth-Values

When there are truth-values of the same judgment from independent resources, NARS integrates them through the following revision rule: $w_+ = w_+^{(1)} + w_+^{(2)}, w_- = w_-^{(1)} + w_-^{(2)}$. $f$ and $c$ are calculated using the above formula.

### A.3.5 Initial Hyperparameter Tunning

Before the experiemnt proper, we use part of the census records from 1899 and earlier as the training data, and the data after that as the testing data. We first sort the census records by individuals, so that records from the same person are arranged together. When using a row of data, we select $m$ more rows after this row to form $m$ pairs. Among these $m$ pairs, some of them are from the same person, which contribute to the positive pattern, and some are from different people, which contribute to the negative pattern.

In testing, we randomly select $x$ individuals in the testing data and use all their related census records to calculate the match score of the records two by two. In classification, we will choose a threshold from 0 to 0.5 and consider that pairs with score > 0.5 + threshold are from the same person, and pairs with score < 0.5 – threshold are from different people, then calculate F1 based on this. The larger the threshold, the more reliable the classification. We discuss the impact of the amount of training data used, the number of patterns used, and the size of the train scope on F1. By default, we use 1/10000 of the training data to train and $x = 1000$ for the testing data to test. The default $m$ is 5. The default number of patterns used is 10.

We found that increasing the number of reference patterns used for recognition improves reliability, while varying the training proportion and observation scope has less consistent effects.

**Handcrafted Rules**: Following the rule-based method, the input of NARS is the summarization of the similarities and differences between two census records. While processing two rows, the following atomic judgments according to different indices will be generated:

***Heimild***: Based on the census year, generate the judgment of “differ in $x$ years”, where $x$ is the absolute value of the difference in heimild in the two rows.

***nafn_norm***: Generate the judgment of “different/same name” based on whether the names are the same. The same applies to ***first_name***, ***patronym***, and ***surname***.

***Birthyear***: Generate the judgment of “different/same birth year” based on whether the birth dates are the same.

***Sex***: Generate the judgment of “different/same gender ”based on whether the genders are the same.

***Status***: It is not enough to only discuss whether the social relations of individuals are the same or not, so two judgments are generated here, namely “the social status is $x$” and “the social status is $y$”, where $x$ and $y$ represent the Status in the two rows of data, in no particular order.

***Marriagestatus***: Generate the judgment of “different/same marriage status” based on whether the marital status is the same.

***Farm***: Generates “different/same farm” judgment based on whether the farm id is the same. The same applies to ***county***, ***parish***, and ***district***.

***Label***: Determines whether the two pieces of data are from the same individual based on whether the person is the same.