INTERSECTIONRE: Mitigating Intersectional Bias in Relation Extraction Through Coverage-Driven Augmentation

Anonymous ACL submission

Abstract

Relation Extraction (RE) models are crucial to many Natural Language Processing (NLP) applications but often inherit and deepen biases in their training data. The underrepresentation of certain demographic groups can result in performance disparities, particularly when considering intersectional fairness, where biases intersect across attributes such as gender and ancestry. To address this issue, we present **INTERSECTIONRE**, a framework to improve the representation of underrepresented groups by generating synthetic training data. **INTERSECTIONRE** identifies gaps in demographic coverage and optimizes data generation, ensuring the quality of augmented data through Large Language Models (LLMs), perplexity scoring, and factual consistency validation. Experimental results on the NYT-10 dataset demonstrate that our approach effectively reduces intersectional disparities and enhances F1 scores, particularly for historically underrepresented groups.

1 Introduction

Relation extraction (RE), a key task in natural language processing (NLP), identifies and classifies semantic relationships between entities (Bunescu and Mooney, 2005). It supports downstream tasks like knowledge graph construction (Muhammad et al., 2020), question-answering (Luo et al., 2020), and information retrieval (Khoo and Myaeng, 2002). Despite strong performance on benchmarks (Cabot and Navigli, 2021; Tang et al., 2022; Orlando et al., 2024), modern neural RE models often exhibit biases across demographic groups (Li et al., 2021; Gaut et al., 2019; Stranisci et al., 2024).

Biases in RE models often stem from their training datasets, directly influencing model predictions (Barocas and Selbst, 2016; Stoyanovich et al., 2020). Poorly curated datasets may underrepresent certain populations due to biased data collection, historical inequalities, or sampling imbalances, leading to discriminatory outcomes and unreliable predictions (Chen et al., 2018; Firmani et al., 2019; Shahbazi et al., 2023). For instance, an RE model trained mostly on data featuring male individuals may struggle with relationships involving female subjects. This systematic underrepresentation, known as *representation bias*, limits the model's ability to generalize across diverse populations (Buolamwini and Gebru, 2018).

Representation bias becomes more complex when multiple demographic attributes intersect, known as *intersectional fairness* (Foulds et al., 2020). Biases can arise within individual groups (e.g., gender or race) and intensify at their intersections. For example, a model may perform well for females and Asians separately but struggle with Asian females due to underrepresentation (Jin et al., 2020). These gaps can lead to systematic RE failures, reinforcing societal biases. Addressing them is crucial for equitable model performance and reducing errors for marginalized groups.

While bias mitigation strategies exist throughout the Machine Learning (ML) pipeline, addressing bias during pre-processing offers a fundamental solution by improving data distribution (Shahbazi et al., 2023). Prior work on analyzing biases in RE, such as (Gaut et al., 2019) revealed gender-based performance disparities, and (Stranisci et al., 2024) expanded analysis to intersectional biases through cross-dataset comparisons. However, they do not propose methods to systematically address intersectional representation gaps.

To address these challenges, we present **INTERSECTIONRE**, a framework for identifying and mitigating intersectional representation gaps in RE datasets. We use pattern-based coverage analysis to quantify demographic representation and identify Maximal Uncovered Patterns (MUPs) to highlight key coverage gaps. We then apply an Integer Linear Programming (ILP) component to determine the minimal number of synthetic examples needed

for balance. Finally, we generate high-quality synthetic data using an LLM-based generator, preserving data characteristics and feature distributions. This approach allows us to balance demographic representation across dimensions while maintaining data integrity.

This study makes three key contributions: (1) **INTERSECTIONRE**, a framework that detects and mitigates intersectional representation gaps in RE tasks through pattern-based gap analysis and synthetic data generation; (2) An ILP-based strategy and LLM-based synthetic data generator to enhance demographic representation while preserving data integrity; (3) Empirical evidence on the NYT-10 dataset (Riedel et al., 2010) showing effective bias mitigation and improved model performance across demographic groups.

2 Background

ML models trained on biased datasets can amplify societal inequalities through unfair predictions (Suresh and Guttag, 2021). In RE models, biased data often results in missing relationships for underrepresented groups. This section examines RE biases, their impacts, and introduces ways to quantify representation and address dataset gaps.

2.1 Relation Extraction and Patterns

Relation Extraction (RE) identifies and classifies semantic relationships between entities in text. Given a sentence x with subject s and object o, the goal is to predict their relation label $y \in \mathcal{Y}$, where $\mathcal{Y}$ is a set of predefined relation types (e.g., `founder`, `employer`). For example, in $x = \{\texttt{Steve Jobs is the founder of Apple}\}$, with $s = \{\texttt{Steve Jobs}\}$ and $o = \{\texttt{Apple}\}$, an RE model identifies $y = \{\texttt{founder}\}$.

A *pattern* P represents a subgroup of records sharing specific attribute values (Asudeh et al., 2019). Formally, P is a vector of size d (number of attributes), where each element $P[i]$ is either a specific value from attribute i's domain or an unspecified value denoted as X. For example, in a dataset with three binary attributes $\{x_1, x_2, x_3\}$, the pattern $P = X01$ includes records with $x_2 = 0$, $x_3 = 1$, and any value for x_1. A record t *matches* pattern P (denoted as $Match(t, P)$) if for all i where $P[i] \neq X$, $t[i] = P[i]$.

To measure representation bias, we use *coverage* to quantify subgroup representation in a dataset $\mathcal{D}$: $Cov(P) = |\{t \in \mathcal{D} \mid Match(t, P)\}|/|\mathcal{D}|$. For example, if $|\mathcal{D}| = 100$ and 21 records match pattern $P = X01$, then $Cov(P) = 0.21$. A pattern P is *uncovered* if $Cov(P) < \tau$, where τ is the minimum required coverage.

Coverage gaps occur when patterns in a dataset are uncovered, leading to potential biases and unfair predictions for these subgroups. Given a dataset $\mathcal{D}$ and a coverage threshold τ, the coverage gap for a pattern P is: $Gap(P) = (\tau - Cov(P)) \times |\mathcal{D}|)$. This represents the minimum number of additional records needed to meet the threshold. For example, if $|\mathcal{D}| = 100$, $Cov(P) = 0.21$, and $\tau = 0.3$, the gap is $(0.3 - 0.21) \times 100 = 9$, meaning nine more records are needed for adequate representation of P.

Two patterns are related through a *parent-child* relationship based on their specified attributes. Pattern P_1 is a *parent* of P_2 ($P_1 \in parent(P_2)$) if it can be formed by replacing exactly one specified value in P_2 with X. Conversely, P_2 is a *child* of P_1 ($P_2 \in child(P_1)$). A pattern can have multiple parents and children. For example, for $P = 101$, its parents are $parent(P) = \{X01, 1X1, 10X\}$, each created by replacing one value with X.

In analyzing coverage gaps, we identify the most general uncovered patterns, called *Maximal Uncovered Patterns (MUPs)*. A pattern P is a MUP if: (1) it is uncovered ($Cov(P) < \tau$) and (2) all its parents have adequate coverage ($\forall P' \in parent(P) : Cov(P') \geq \tau$). MUPs capture broad underrepresented subgroups without redundancy from more specific child patterns.

2.2 Problem Definition

Given a RE dataset $\mathcal{D}$ with triples (subject s, relation r, object o) and demographic attributes (gender $\mathcal{G}$, ancestry $\mathcal{A}$), the goal is to mitigate biases from coverage gaps, especially intersectional ones, that affect model performance for underrepresented groups. We analyze intersectional representation using patterns P and identify MUPs to address gaps without redundant subpattern analysis.

Improving MUP coverage is crucial because MUPs represent the broadest underrepresented subgroups, and by increasing coverage for these general patterns, we automatically improve the coverage of all their more specific child patterns. For each MUP M, at least $Gap(M)$ additional records are needed to meet the coverage threshold τ. This process balances fairness across gender and ancestry while minimizing synthetic data to preserve data quality.

2.3 Synthetic Data Generation

Synthetic data generation is a key approach to addressing representation bias in ML datasets, where imbalanced demographics can lead to discriminatory model behavior (Draghi et al., 2021). It helps mitigate biased predictions by balancing demographic attributes while minimizing generated records (Wang et al., 2024). However, balancing representation is challenging, especially with intersectional attributes (Shahbazi et al., 2023), due to the difficulty of ensuring proportional representation across dimensions (e.g., gender, race) while addressing coverage gaps (Fournier-Montgieux et al., 2024). For example, if Black females are underrepresented compared to Black males or Asian females, data generation must fill this gap without disrupting other balances. Overcompensation can create new biases, making it hard to maintain fairness and data integrity.

To address these challenges, we optimize synthetic data generation to minimize records while meeting representation goals (Micheletti et al., 2023). Traditional greedy algorithms often yield suboptimal results and struggle to maintain demographic balance (Shahbazi et al., 2024; Erfanian et al., 2024). To overcome this, we use Integer Linear Programming (ILP) (Nandwani et al., 2022) to define coverage and balance constraints, minimizing synthetic records while ensuring fair representation across all intersections (Dwork et al., 2024). This approach is especially effective for MUPs, providing globally optimal solutions that satisfy all gaps and constraints. The next section details our ILP formulation and implementation.

3 INTERSECTIONRE

This section presents **INTERSECTIONRE** for addressing intersectional representation bias in RE datasets, consisting of five components: (1) a data enrichment pipeline adding demographic attributes, (2) a pattern identification algorithm detecting underrepresented groups via MUP analysis, (3) an ILP-based planner minimizing required records while ensuring balance, (4) an entity collection module sourcing data from knowledge bases, and (5) an LLM-based generator producing synthetic factual samples. The following sections detail each component's role in mitigating bias.

3.1 Data Enrichment Pipeline

Analyzing intersectional fairness in RE datasets requires demographics (e.g., gender, ancestry), which are often missing (Stranisci et al., 2024). For example, a record like (`Steve Jobs`, `Founder`, `Apple`) lacks demographic details.

To address this, we developed a data enrichment pipeline using Wikidata to extract demographic attributes, consisting of two stages: First, for each record, we focus on relation labels, such as `founder`, `place_of_birth`, `profession`, and `nationality` that involve human entities, excluding records without them to ensure relevant demographic analysis. Then, for identified human entities, we retrieve attributes like gender and citizenship from Wikidata. We map each country to a broader ancestry group (e.g., `African`, `Asian`, `European/Western`, `Latino/Caribbean`, `Middle Eastern`) using a curated country-to-ancestry mapping, enabling meaningful aggregation to identify representation patterns and coverage gaps.

3.2 Pattern Identification

After enriching the dataset with demographic attributes, we identify underrepresented groups by analyzing coverage patterns based on gender and ancestry, focusing on MUPs that represent the broadest coverage gaps. Since identifying all MUPs is computationally intensive (Shahbazi et al., 2023), we propose an algorithm inspired by DEEPDIVER (Asudeh et al., 2019). DEEPDIVER uses a hybrid strategy combining downward traversal with immediate upward verification, checking all ancestor patterns to confirm MUPs, but our approach simplifies this by verifying only the immediate parent during downward traversal and deferring full maximality checks to a post-processing phase. We apply two pruning strategies: (1) *Coverage-Based Pruning*, where patterns meeting or exceeding the threshold have their children explored as potential MUPs, and (2) *Parent-Based Pruning*, where patterns below the threshold are pruned if their immediate parent is also uncovered. This reduces verification overhead, with post-processing ensuring only maximal patterns are retained.

As shown in Figure 1, consider a dataset with attributes Gender: `Male` (M), `Female` (F)} and Ancestry: {`Asian` (A), `European` (E), `Latino` (L)}, and a threshold $\tau = 0.3$. Starting from the root XX (coverage 1.0), its children MX and FX are explored since XX exceeds the threshold. MX (coverage 0.8) is not a MUP, so its children MA, ME, and ML are explored. FX (coverage 0.2) is a potential MUP, and *Coverage-Based Pruning* skips its children (FA, FE, FL) since their parent

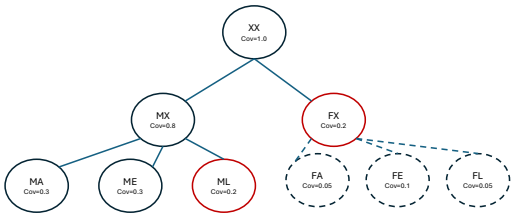


Figure 1: Tree structure illustrating DFS-based MUP discovery with pruning. Red nodes represent identified MUPs, while dashed nodes and edges indicate pruned patterns and paths.

is already uncovered. For ML (coverage 0.2), our algorithm checks only its immediate parent (MX), unlike DEEPDIVER, which checks both MX and XX. This streamlined approach flags ML as a potential MUP, with maximality verified during post-processing.

3.3 ILP-based Generation Plan

After identifying MUPs, we propose an ILP-based planning component to minimize synthetic records while meeting coverage requirements. Unlike greedy algorithms (Erfanian et al., 2024), which require iterative MUP recalculations and struggle to maintain demographic balance, our ILP ensures global optimality in a single step. It minimizes synthetic records under two constraints: (1) generating at least $Gap(M)$ records per MUP to meet coverage thresholds and (2) maintaining balanced gender ratios within each ancestry group. This prevents addressing gaps for one group (e.g., Asian females) from creating imbalances in others.

Let $\mathcal{G}$ = {`Female`, `Male`} and $\mathcal{A}$ = {`African`, `Asian`, `European/Western`, `Latino/Caribbean`, `Middle Eastern`}. To avoid new biases, we track for each ancestry $a \in \mathcal{A}$ the number of female records (F_a), total records (T_a), and female ratio ($R_a = F_a/T_a$). Simply adding new records can skew the balance. For example, for MUP $M_1 = \{\texttt{Female}, \texttt{Asian}\}$ with 0.01 coverage in a dataset of 1000 records and threshold $\tau = 0.05$ where the pattern $P = \{\texttt{X}, \texttt{Asian}\}$ has the coverage of 0.05, the gap $Gap(M_1) = 40$ requires 40 more records. Adding only female records would skew gender balance, so the ILP determines how many male Asian records to add to maintain fairness. To formulate this as an ILP, we define decision variables ($x_{g,a} \geq 0, \forall g \in \mathcal{G}, a \in \mathcal{A}$) indicating the number of synthetic records to generate for each gender and ancestry combination. These variables are only active for demographic combinations linked to MUPs, minimizing unnecessary data generation.

The next step in the ILP formulation is defining the objective function. Our primary goal is to minimize the total number of synthetic records required to meet demographic coverage and intersectional balance requirements: minimize $\sum_{g\in\mathcal{G}} \sum_{a\in\mathcal{A}} x_{g,a}$. This minimization ensures efficient data generation by creating only the necessary records to address coverage gaps identified by MUPs.

Then, we need to define the constraints. Our ILP formulation includes two constraints to ensure adequate coverage and intersectional balance: (1) *coverage constraints* and (2) *gender balance constraints*. To satisfy the *coverage constraints*, for each MUP ($M \in \mathcal{M}$), we ensure coverage gaps are filled: $\sum_{g\in\mathcal{G}_M} \sum_{a\in\mathcal{A}_M} x_{g,a} \geq Gap(M)$, where $\mathcal{G}_M$ and $\mathcal{A}_M$ represent the gender and ancestry sets specified in MUP M. For specified attributes (e.g., `Female`), the set contains only that value, and for unspecified attributes (X), it includes all possible values.

To satisfy the *gender balance constraints*, for each ancestry group $a \in \mathcal{A}_M$, we implement adaptive gender balance constraints:

$$\text{min\_}R_a \leq \frac{F_a + x_{\text{female},a}}{T_a + x_{\text{female},a} + x_{\text{male},a}} \leq \text{max\_}R_a, \quad (1)$$

where F_a and T_a represent the current number of female and total records in group a. The variables $x_{\text{female},a}$ and $x_{\text{male},a}$ are decision variables for generating female and male records in ancestry group a. The bounds adapt based on the current ratio (R_a) and the severity of the imbalance:

$$min\_R_a = \begin{cases} \min(\alpha_1 \times R_a, 0.5) & \text{if } R_a < 0.33 \text{ (severe)} \\ \min(\alpha_2 \times R_a, 0.45) & \text{otherwise} \end{cases} \quad (2)$$

$$max\_R_a = \begin{cases} \max(\beta_1 \times R_a, 0.5) & \text{if } R_a < 0.33 \text{ (severe)} \\ \max(\beta_2 \times R_a, 0.55) & \text{otherwise} \end{cases} \quad (3)$$

The threshold of 0.33 reflects a 2:1 male-to-female ratio based on fairness literature (Stranisci et al., 2024). Parameters α_1, α_2, β_1, and β_2 control adjustment rates for severe and moderate imbalances. The ILP output is a generation plan with decision variables ($x_{g,a}, \forall g \in \mathcal{G}_M, a \in \mathcal{A}_M$) indicating the minimal records needed for each intersectional group to close coverage gaps.

3.4 Entity Collection from Knowledge Bases

To generate realistic records, we convert the ILP-based plan into synthetic data using Wikidata (structured) and Wikipedia (unstructured). For each gender–ancestry pair, we map ancestry to countries

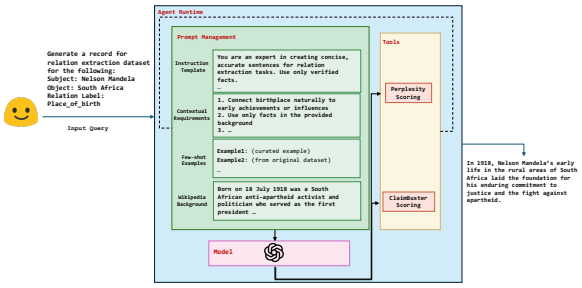


Figure 2: LLM-based record generation architecture, featuring prompt management providing relation-specific prompts for the model, with generated sentences validated by Perplexity Scoring and ClaimBuster tools.

(Section 3.1), distribute entities accordingly, and apply per-citizenship limits for diversity. SPARQL queries retrieve Wikidata entities with matching gender and citizenship, along with biographical details (e.g., `founder`, `employer`, `place_lived`, `religion`, `profession`, `nationality`). To enrich context, we also fetch Wikipedia introductions via the MediaWiki API. This blend of structured and unstructured data ensures factual accuracy while meeting demographic requirements.

3.5 LLM-based Record Generation

Our framework's final stage uses a GPT-4-powered AI agent to generate synthetic records. Figure 2 illustrates the agent's architecture, which consists of prompt management components, a core generation model, and validation tools. The agent uses GPT-4 to map each relation-specific prompt p to synthetic records x. Each prompt p is tailored to capture the unique traits of a relation label $y \in \mathcal{Y}$, ensuring the generated sentence x accurately reflects the entity relationship. In this process, the agent addresses two validation challenges: (1) *distribution alignment*, ensuring x matches the linguistic and structural patterns of the original dataset $\mathcal{D}$, and (2) *factual consistency*, ensuring x accurately reflects input relationships. It uses a Perplexity Scoring Tool for language alignment and ClaimBuster (Jimenez and Li, 2018) for factual consistency.

To guide GPT-4, we design relation-specific prompts p with the following components (Figure 2): (1) a *system prompt & instruction template* tailored to each relation y, defining constraints and guidelines, (2) *contextual requirements*, focusing on verified facts, achievements, or relevance (e.g., `lived_in` for locations, `employer` for roles), and (3) *few-shot examples*, combining curated and dynamic samples for diverse, in-context guidance.

The agent incorporates two key validation mechanisms to ensure the quality of generated records: *distribution alignment* and *factual consistency*. For distribution alignment, we measure perplexity per relation using a pre-trained model (e.g., GPT-2), where lower perplexity indicates better fluency and alignment. Specifically, we ensure that the perplexity of any generated sentence does not exceed the mean plus two standard deviations of perplexity values calculated for existing sentences of the same relation label. This method confirms that generated sentences maintain a consistent quality and style with the dataset's typical variability.

For factual consistency, the agent uses ClaimBuster with dynamic thresholding. Let $\phi(x)$ be the ClaimBuster scoring function assigning a factuality score in [0,1]. For each relation y, we set the threshold θ_r at the 25th percentile of original dataset scores: $\theta_r = \text{percentile}_{25}(\{\phi(x) \mid x \in \mathcal{D}, \text{relation}(x) = y\})$, ensuring generated sentences are at least as factual as 75% of the original dataset. Sentences must meet $\phi(x) \geq \theta_r$; those below are refined with stricter prompts and re-evaluated. Only sentences passing after either stage are accepted, ensuring high factual consistency.

The agent iteratively refines and regenerates sentences using adjusted prompts until they meet both distributional and factual standards or reach a retry limit, ensuring the generation of high-quality, realistic synthetic records that effectively address representation gaps.

4 Experimental Results and Analysis

This section focuses on evaluating our framework for addressing intersectional representation bias in relation extraction datasets, specifically: (1) improving demographic representation, (2) efficient synthetic data generation via ILP, and (3) impact on model performance across subgroups.

4.1 Experimental Setup

Dataset. We conduct our experiments on the NYT-10 dataset (Riedel et al., 2010), a benchmark for RE tasks with 70,339 records and 52 relation labels from the New York Times corpus, annotated via distant supervision from Freebase. To enable demographic analysis, we filtered for records with at least one human entity, yielding 30,818 records and 15 human-centric relation labels, with the most frequent being `place_of_birth` (21.3%), `nationality` (18.7%), `employer` (15.4%), and `place_lived` (14.2%). We enriched the dataset

with demographic attributes like gender and ancestry (Section 3.1), revealing a gender imbalance (12.4% `females` vs. 87.6% `males`) and ancestry disparities (`European/Western` 71.1%, `Middle Eastern` 11.7%, `Asian` 9.2%, `Latino/Caribbean` 4.9%, `African` 3.1%). These imbalances highlight the need to address intersectional coverage gaps for equitable representation.

Implementation. We queried demographic attributes from Wikidata using SPARQL, optimized via `SPARQLWrapper`, and pre-designed citizenship to ancestry mappings. The ILP was formulated using Gurobi, applying dynamic gender balance constraints based on R_a (stricter when $R_a < 0.33$: $\alpha_1 = 1.5, \beta_1 = 2$; relaxed otherwise: $\alpha_2 = 0.9, \beta_2 = 1.1$). Synthetic records were generated with GPT-4 (200-token limit, temperature 0.0) and validated via GPT-2 perplexity scoring (Radford et al., 2019) for fluency and ClaimBuster (Jimenez and Li, 2018) for factual consistency. We fine-tuned the REBEL-large model (Cabot and Navigli, 2021) (a seq2seq BART-based RE model (Lewis, 2019)) on NYT-10, training for 3 epochs with AdamW (learning rate $2e-4$, batch size 4).

4.2 Intersectional Representation Analysis

We analyzed intersectional representation patterns in the original dataset and our proposed solutions. To evaluate our ILP-based constraints, we conducted experiments in three settings: (1) the original dataset, (2) our framework with intersectional balance constraints, and (3) our framework with constraints off, focusing on the MUP coverage threshold.

Basleine Coverage Gap in Original Dataset. To assess intersectional gaps, we used a coverage threshold of 0.15, representing the minimal expected coverage per group, given five ancestry groups and two genders (ideally 10% each if evenly distributed). This value balances real-world demographic imbalances with meaningful targets for underrepresented groups. Figure 3 (Left) shows demographic representation in the original dataset, with rectangles indicating the percentage of each gender-ancestry combination and color intensity reflecting coverage (darker = higher). `European/Western` males dominate (61.7%), while female representation is minimal, peaking at 9.4% and dropping to 0.3% for African females. Groups like `African males` (2.8%) and `Latino/Caribbean females` (0.4%) fall well below the threshold, highlighting systemic biases and the need for targeted augmentation.

Coverage Improvements with Augmentation. Figure 3 also shows the impact of our augmentation strategies in two scenarios: *With Intersectional Balance Constraint* (middle) and *Without Intersectional Balance Constraint* (right). Each rectangle reflects adjusted percentages after augmentation. With the constraint, demographic representation becomes more equitable, increasing female representation to 7.8% across ancestries and helping most groups reach the 0.15 threshold. This approach effectively addresses under-representation (e.g., `African` and `Latino/Caribbean` females) while maintaining proportionality, ensuring balanced improvements without new biases. The *Without Intersectional Constraint* scenario results are uneven. Some underrepresented groups improve, but `European/Western males` disproportionately benefit, rising to 49.4%, while groups like `Middle Eastern` and `Asian` females remain below the 0.15 threshold. This highlights the need for balance constraints to achieve fair coverage.

Gender Ratios Across Approaches. To evaluate gender equity across ancestry groups, we computed the female-to-male ratio for each group. An ideal ratio of 0.5 indicates equal representation, yet in the original dataset, ratios are skewed, with females comprising less than 0.15 in most groups, showing severe under-representation. Applying intersectional balance constraints achieves near-parity across ancestries, effectively addressing these imbalances—for example, `African` and `Middle Eastern` groups reach ratios close to 0.5 from near-zero. In contrast, the absent of such constraints leads to partial improvements but fails to ensure consistent gender equity. These results highlight the necessity of intersectional balance constraints for equitable representation.

Intersectional Coverage vs. Synthetic Records Trade-Off. Figure 4a shows a Pareto analysis of the trade-off between intersectional coverage improvement and the number of synthetic records added. The *With Intersectional Constraint* strategy achieves the highest improvement (0.076) with 27,913 records, balancing fairness and efficiency. In contrast, the *Without Intersectional Constraint* strategy shows lower improvement (0.039) with 7,668 records, highlighting its inefficiency in addressing intersectional gaps. The original dataset serves as the baseline with no synthetic records or

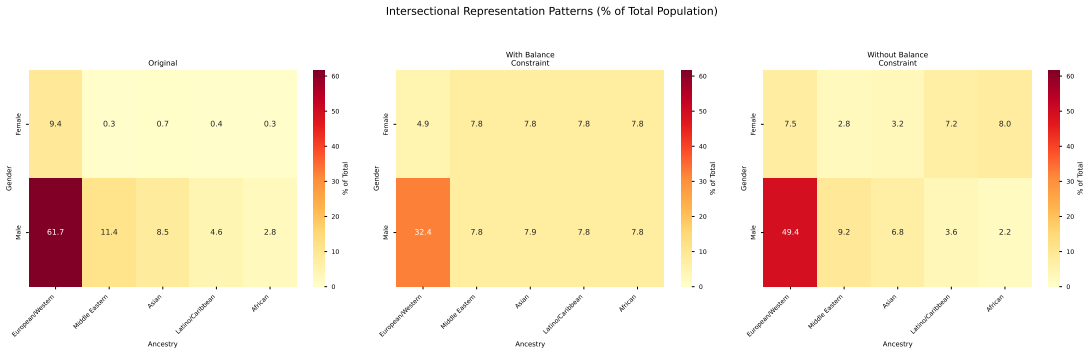


Figure 3: Intersectional representation across the original dataset, generated dataset with intersectional constraint, and generated dataset without intersectional constraint.

improvement. While requiring more records, the constrained strategy achieves significantly greater coverage improvements than the unconstrained approach.

Intersectional Fairness Metrics. To evaluate fairness across the three dataset states, we analyzed four normalized metrics ($[0, 1]$), where higher values indicate better representation: (1) *Balance Score* (normalized female-to-male ratio), (2) *Gender Gap* (difference in female and male representation), (3) *Ancestry Gap* (standard deviation across ancestry groups), and (4) *Intersectional Gap* (combined gender and ancestry disparities). Figure 4b shows these metrics. The original dataset shows significant disparities with consistently low scores (≤ 0.15) across all metrics, while the *Without Intersectional Constraints* approach shows moderate, uneven improvements (0.35–0.45). In contrast, our *With Intersectional Constraints* approach achieves the highest scores, notably for ancestry (0.75) and intersectional gaps (0.65), effectively mitigating representation biases. The Balance Score and Gender Gap improve from 0.124 (original) to 0.569 (constrained), reducing gender disparities while maintaining ancestry balance.

Statistical Consistency. We compared sentence length distributions between the original NYT-10 and the augmented dataset to assess stylistic consistency. Figure 4c shows closely aligned density curves, supported by a low Jensen-Shannon Divergence (0.0411) and KS test statistic (0.0491, $p < 0.0001$). Sentence length statistics confirm this: the original dataset has a mean of 40.95, a median of 39.00, and a standard deviation (SD) of 78.92, while the augmented dataset shows a mean of 39.81, a median of 37.00, and an SD of 75.55, indicating minimal deviation.

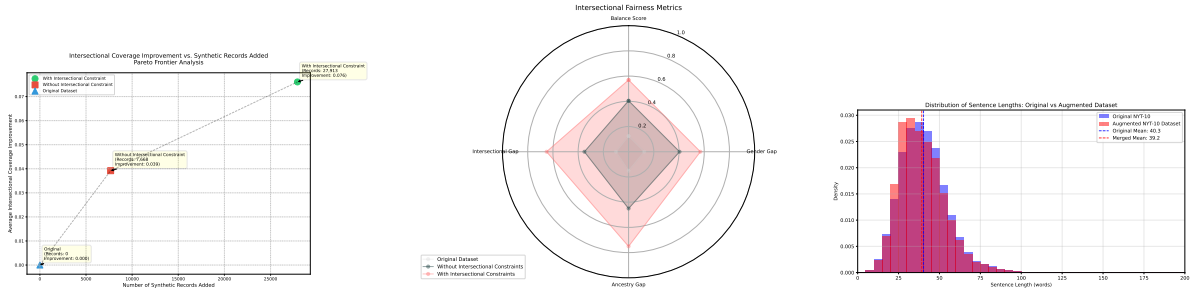
For quality assessment, the vocabulary size grew from $37,168$ to $42,862$, showing that the augmented dataset introduces new vocabulary while maintaining a reasonable growth rate. This suggests the generated text preserves the domain-specific language of the original dataset. The *Type-Token Ratio* (TTR), measuring lexical diversity as the ratio of unique words to total words, rose slightly from 0.0349 to 0.0378 (+8.3%), maintaining diversity without excessive repetition. The *Hapax Percentage*, indicating the proportion of words appearing only once, increased from 24.71% to 27.60% (+11.7%), reflecting more unique terms, likely from new entity names. These results demonstrate that our augmentation approach effectively enhances coverage and diversity while preserving linguistic and structural integrity.

Table 1: Model Performance Comparison Across Demographic Groups

		Original		Augmented		Δ F1
Gender	Ancestry	F1 Score	FPR	F1 Score	FPR	Disparity
Overall	—	0.782	0.197	0.845	0.265	+0.063
Female	African	0.000	1.000	1.000	0.000	+1.000
	Asian	0.773	0.074	0.941	0.111	+0.168
	European/Western	0.795	0.199	0.890	0.199	+0.095
	Latino/Caribbean	0.889	0.200	1.000	0.000	+0.111
	Middle Eastern	0.870	0.020	0.950	0.015	+0.080
Male	African	0.756	0.179	0.923	0.143	+0.167
	Asian	0.902	0.178	0.861	0.200	-0.041
	European/Western	0.805	0.323	0.755	0.228	-0.050
	Latino/Caribbean	0.911	0.116	0.911	0.163	+0.000
	Middle Eastern	0.890	0.025	0.950	0.018	+0.060

4.3 Model Performance

Table 1 shows significant variations in the REBEL model's performance when fine-tuned on the original NYT-10 dataset versus the demographically augmented version. The augmented model's F1 score improves from 0.782 to 0.845, reflecting better overall performance, though gains are uneven across demographic groups. Notably,



(a) (b) (c)

Figure 4: Analysis of representation improvements across different augmentation strategies. Subfigure (a) shows the trade-off between coverage improvement and synthetic records added, while subfigure (b) compares intersectional fairness metrics across dataset states. Subfigure (c) shows the sentence length comparison between the Original NYT-10 and Augmented Dataset. Statistical tests confirm a high degree of alignment, with minor deviations in mean and variance.

underrepresented groups like `African males`, `African females`, `Middle Eastern females`, and `Latino/Caribbean females` see substantial improvements, indicating the augmentation effectively addresses representation gaps. While majority groups such as `European/Western males` show a slight F1 decrease (-0.050), this is offset by minority group gains. The augmented model also reduces false positive rates (FPR) across most demographics while maintaining strong performance for Middle Eastern groups. However, these results are influenced by demographic imbalances in the test set, potentially affecting metric reliability for underrepresented groups. This highlights the need for evaluation methods that consider representation in both the training and testing phases.

5 Related Work

Bias in RE has been widely studied, especially regarding gender disparities. Gender-based biases have received particular attention, with WikiGenderBias (Gaut et al., 2019) revealing significant performance disparities in occupation and spouse-related relations. Similarly, (Stranisci et al., 2024) demonstrated that RE datasets systematically underrepresent non-Western nationalities and female entities, leading to biased model behavior. Entity-level biases represent another critical challenge in RE systems. (Wang et al., 2022) showed that RE models disproportionately rely on entity mentions rather than contextual information, proposing counterfactual inference as a mitigation strategy at inference time. Building on this work, (He et al., 2025) developed DREB, a debiased benchmark that addresses entity bias through systematic entity replacement, and introduced MixDebias, which combines data augmentation with model-level debiasing. However, while their approach effectively reduces entity bias, their entity replacement strategy can generate factually incorrect relationships and does not address underlying demographic representation gaps in training data. The quality and fairness of training data itself have also been investigated. (Li et al., 2020) identified systematic biases in distantly supervised datasets, noting that conventional held-out evaluations may misrepresent model fairness due to label noise. Unlike previous work focusing on bias detection or implementing mitigation strategies at the cost of factual correctness, our approach proactively addresses bias at the data level through coverage-driven augmentation, generating synthetic data to create balanced, fairer RE datasets while maintaining factual accuracy.

6 Conclusion

This work tackles intersectional fairness in relation extraction (RE) datasets, addressing representation bias that leads to disproportionate model errors for underrepresented groups. We propose **INTERSECTIONRE** to identify and mitigate demographic coverage gaps, ensuring balanced representation across gender and ancestry while preserving linguistic and factual integrity. Empirical results show that our augmentation strategy improves demographic representation, reduces disparities, and enhances the REBEL model's F1 score, especially for underrepresented groups. Our findings demonstrate the effectiveness of structured augmentation in mitigating demographic bias. Future work should extend this framework to include more attributes (e.g., age, profession), diversify demographic sources beyond Wikidata, and move beyond binary gender classifications. Our approach offers a scalable, adaptable method for promoting demographic fairness in RE, supporting more equitable AI systems.

7 Limitations

While this study demonstrates the effectiveness of **INTERSECTIONRE** in mitigating intersectional bias in RE, several limitations should be acknowledged.

First, our framework relies on external knowledge bases (e.g., Wikidata) for demographic annotations. While these sources offer extensive coverage, may contain gaps or inaccuracies, particularly for individuals from less-documented regions or historical contexts. The quality of demographic inference directly impacts the effectiveness of our augmentation strategy. The errors in entity annotation could propagate through the dataset. Future research directions could investigate more sophisticated demographic inference techniques, including human-in-the-loop validation mechanisms, to improve the robustness and reliability of the annotation process.

Second, the current implementation models gender as a binary attribute (male/female) due to constraints in demographic annotations. This oversimplifies real-world gender diversity and may reinforce binary assumptions in NLP models. Future extensions should explore more inclusive demographic attributes, including non-binary and genderfluid identities, to ensure broader fairness.

Third, the synthetic data generation process using generative AI introduces substantial computational and financial costs. The generation of high-quality synthetic data requires significant computational resources, while API access to advanced LLMs presents cost barriers. Future research could explore more efficient alternatives, such as lightweight models or Retrieval-Augmented Generation (RAG) techniques, to reduce dependence on LLMs while maintaining data quality.

These limitations suggest directions for future research, including a deeper exploration of knowledge bases, improved demographic annotation strategies and cost-efficient synthetic data generation methods.

References

Abolfazl Asudeh, Zhongjun Jin, and HV Jagadish. 2019. Assessing and remedying coverage for a given dataset. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 554–565. IEEE.

Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.*, 104:671.

Razvan Bunescu and Raymond Mooney. 2005. A shortest path dependency kernel for relation extraction. In *EMNLP*, pages 724–731.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. Rebel: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381.

Irene Chen, Fredrik D Johansson, and David Sontag. 2018. Why is my classifier discriminatory? *Advances in neural information processing systems*, 31.

Barbara Draghi, Zhenchen Wang, Puja Myles, and Allan Tucker. 2021. Bayesboost: Identifying and handling bias using synthetic data generators. In *Third International Workshop on Learning with Imbalanced Domains: Theory and Applications*, pages 49–62. PMLR.

Cynthia Dwork, Kristjan Greenewald, and Manish Raghavan. 2024. Synthetic census data generation via multidimensional multiset sum. *arXiv preprint arXiv:2404.10095*.

Mahdi Erfanian, H. V. Jagadish, and Abolfazl Asudeh. 2024. Chameleon: Foundation models for fairness-aware multi-modal data augmentation to enhance coverage of minorities. *Proc. VLDB Endow.*, 17(11):3470–3483.

Donatella Firmani, Letizia Tanca, and Riccardo Torlone. 2019. Ethical dimensions for data quality. *Journal of Data and Information Quality (JDIQ)*, 12(1):1–5.

James R Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2020. Bayesian modeling of intersectional fairness: The variance of bias. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, pages 424–432. SIAM.

Alexandre Fournier-Montgieux, Michael Soumm, Adrian Popescu, Bertrand Luvison, and Hervé Le Borgne. 2024. Fairer analysis and demographically balanced face generation for fairer face verification. *arXiv preprint arXiv:2412.03349*.

Andrew Gaut, Tony Sun, Shirlyn Tang, Yuxin Huang, Jing Qian, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, et al. 2019. Towards understanding gender bias in relation extraction. *arXiv preprint arXiv:1911.03642*.

Liang He, Yougang Chu, Zhen Wu, Jianbing Zhang, Xinyu Dai, and Jiajun Chen. 2025. Rethinking relation extraction: Beyond shortcuts to generalization with a debiased benchmark. *arXiv preprint arXiv:2501.01349*.

Damian Jimenez and Chengkai Li. 2018. An empirical study on identifying sentences with salient factual statements. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Zhongjun Jin, Mengjing Xu, Chenkai Sun, Abolfazl Asudeh, and HV Jagadish. 2020. Mithracoverage: a system for investigating population bias for intersectional fairness. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pages 2721–2724.

Christopher Khoo and Sung Hyon Myaeng. 2002. Identifying semantic relations in text for information retrieval and information extraction. In *The semantics of relationships: An interdisciplinary perspective*, pages 161–180.

Mike Lewis. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Luoqiu Li, Xiang Chen, Hongbin Ye, Zhen Bi, Shumin Deng, Ningyu Zhang, and Huajun Chen. 2021. On robustness and bias analysis of bert-based relation extraction. In *Knowledge Graph and Semantic Computing: Knowledge Graph Empowers New Infrastructure Construction: 6th China Conference, CCKS 2021, Guangzhou, China, November 4-7, 2021, Proceedings 6*, pages 43–59. Springer.

Pengshuai Li, Xinsong Zhang, Weijia Jia, and Wei Zhao. 2020. Active testing: An unbiased evaluation method for distantly supervised relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 204–211.

Da Luo, Jindian Su, and Shanshan Yu. 2020. A bert-based approach with relation-aware attention for knowledge base question answering. In *2020 IJCNN*. IEEE.

Nicolo Micheletti, Raffaele Marchesi, Nicholas I-Hsien Kuo, Sebastiano Barbieri, Giuseppe Jurman, and Venet Osmani. 2023. Generative ai mitigates representation bias and improves model fairness through synthetic health data. *medRxiv*, pages 2023–09.

Iqra Muhammad, Anna Kearney, et al. 2020. Open information extraction for knowledge graph construction. In *DEXA*, pages 103–113.

Yatin Nandwani, Rishabh Ranjan, Parag Singla, et al. 2022. A solver-free framework for scalable learning in neural ilp architectures. *Advances in Neural Information Processing Systems*, 35:7972–7986.

Riccardo Orlando, Pere-Lluís Huguet Cabot, Edoardo Barba, and Roberto Navigli. 2024. Relik: Retrieve and link, fast and accurate entity linking and relation extraction on an academic budget. *arXiv preprint arXiv:2408.00103*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part III 21*, pages 148–163. Springer.

Nima Shahbazi, Mahdi Erfanian, and Abolfazl Asudeh. 2024. Coverage-based data-centric approaches for responsible and trustworthy ai. *IEEE Data Eng. Bull.*, 47(1):3–17.

Nima Shahbazi, Yin Lin, Abolfazl Asudeh, and HV Jagadish. 2023. Representation bias in data: A survey on identification and resolution techniques. *ACM Computing Surveys*, 55(13s):1–39.

Julia Stoyanovich, Bill Howe, and Hosagrahar Visvesvaraya Jagadish. 2020. Responsible data management. *Proceedings of the VLDB Endowment*, 13(12).

Marco Stranisci, Pere-Lluís Huguet Cabot, Elisa Bassignana, and Roberto Navigli. 2024. Dissecting biases in relation extraction: A cross-dataset analysis on people's gender and origin. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 190–202, Bangkok, Thailand. Association for Computational Linguistics.

Harini Suresh and John Guttag. 2021. A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–9.

Wei Tang, Benfeng Xu, Yuyue Zhao, Zhendong Mao, Yifeng Liu, Yong Liao, and Haiyong Xie. 2022. Unirel: Unified representation and interaction for joint relational triple extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7087–7099.

Ke Wang, Jiahui Zhu, Minjie Ren, Zeming Liu, Shiwei Li, Zongye Zhang, Chenkai Zhang, Xiaoyu Wu, Qiqi Zhan, Qingjie Liu, et al. 2024. A survey on data synthesis and augmentation for large language models. *arXiv preprint arXiv:2410.12896*.

Yiwei Wang, Muhao Chen, Wenxuan Zhou, Yujun Cai, Yuxuan Liang, Dayiheng Liu, Baosong Yang, Juncheng Liu, and Bryan Hooi. 2022. Should we rely on entity mentions for relation extraction? debiasing relation extraction with counterfactual analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3071–3081.

A Appendix: Methodology Overview

To provide a clearer understanding of our approach, we include an overview of our methodology in Figure 5. Our pipeline consists of five main stages:

- **Enrichment Pipeline**: Extracting demographic attributes from Wikidata and filtering relations relevant for augmentation.
- **Pattern Identification**: Identifying MUPs and analyzing their coverage.
- **ILP-Based Planning**: Formulating an ILP model to calculate a generation plan to balance demographic representation in the dataset.
- **Entity Collection**: Retrieving entity details from Wikidata using SPARQL queries based on the generation plan.
- **LLM-Based Generation**: Generating synthetic data using GPT-4, with validation tools ensuring factual accuracy and linguistic fluency.

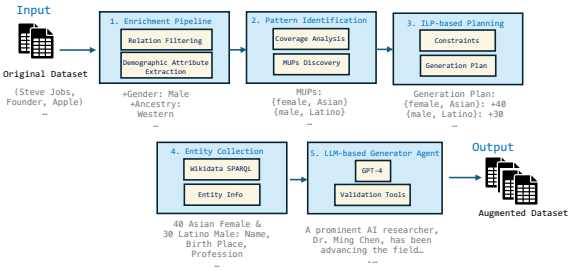


Figure 5: Overview of the data augmentation methodology, illustrating the key processing steps from the original dataset to the final augmented dataset.

B Prompt Templates for Synthetic Data Generation

To ensure high-quality and contextually accurate synthetic data generation, we designed **relation-specific prompts** tailored to different relation extraction tasks. The prompt structure consists of four key components:

- **System Prompt**: Defines the model's role and ensures that generated sentences adhere to desired linguistic and factual constraints.
- **Contextual Requirements**: Specifies key constraints and stylistic elements to maintain factual accuracy and fluency.
- **Few-Shot Examples**: Provides real-world examples to guide generation.
- **Entity-Specific Context**: Includes subject, object, relation label, gender, and background information.

Figure 6 illustrates our prompt templates for two representative relation labels: *place_lived* and *place_of_birth*. Each template is carefully structured to guide the generation of natural sentences that implicitly convey the intended relationship. For instance, the *place_lived* template emphasizes connecting locations to significant work achievements, while the *place_of_birth* template focuses on early life influences and cultural context.

You are an expert in crafting concise, natural sentences about where people lived, focusing on verified historical facts. Your task is to create a single sentence that:

- Uses only time periods mentioned in the provided background
- Connects location naturally to a single significant aspect of their work
- Maintains historical accuracy without speculation
- Avoids complex, multi-clause structures
- Never invent or infer information not present in the background
- Creates clear cause-and-effect relationships between location and achievement

Generate one focused sentence that:

1. Uses only time periods explicitly mentioned in the background
2. Highlights one specific achievement or activity from their known history
3. Shows how the location influenced or enabled this achievement
4. Incorporates verified cultural/social elements from their background
5. Keeps the relationship between person and location subtle but clear

Important:

- Focus on one main idea rather than multiple achievements
- Use only facts provided in the background
- Create a clear but natural connection to the location
- Aim for 20-30 words for clarity and impact

Given the following examples from real-world text showing how relations are expressed naturally:

Examples:
Subject: Leonard Bernstein,
Relation label: place_lived,
Object: New York,
Sentence: Throughout the 1960s, many of Bernstein's most innovative compositions took shape in his Upper West Side studio, where the maestro would often host late-night rehearsals with the New York Philharmonic.

Examples from Original Dataset

Now, generate a similar natural sentence for the following relation. The sentence should avoid directly stating the relationship and should sound natural in a relation extraction dataset.

- Subject: subject entity
- Relation: place_lived
- Object: object entity
- Background: entity Wikipedia content

You are an expert in crafting natural sentences about early life and origins. Your task is to create a single sentence that:

- Uses only dates and facts mentioned in the provided background
- Connects birthplace naturally to early achievements or influences
- Maintains historical accuracy without speculation
- Avoids formulaic birth-related phrases
- Never invents or infers information not present in the background

Generate one focused sentence that:

1. Uses specific dates/periods from the background
2. Highlights one early achievement or influence
3. Places birthplace naturally within the narrative
4. Incorporates verified cultural or historical context
5. Keeps the birthplace reference subtle but clear

Important:

- Never use obvious phrases like "was born in"
- Connect location to early life or achievements
- Use only facts provided in the background
- Aim for 20-30 words with natural flow

Given the following examples from real-world text showing how relations are expressed naturally:

Examples:
Subject: Gabriel García Márquez,
Relation label: place_of_birth,
Object: Aracataca,
Sentence: The magical realism in García Márquez's stories drew deep inspiration from his childhood in Aracataca, where his grandmother's storytelling shaped his earliest literary sensibilities.

Examples from Original Dataset

Now, generate a similar natural sentence for the following relation. The sentence should avoid directly stating the relationship and should sound natural in a relation extraction dataset.

- Subject: subject entity
- Relation: place_of_birth
- Object: object entity
- Background: entity Wikipedia content

Figure 6: Prompt templates for generating sentences for place_lived (left) and place_of_birth (right) relation labels. The templates include instruction template (system prompt), context requirements, and few-shot examples to guide the generation of natural sentences.