applied sciences

MDPI

Article

Spatiotemporal Lattice-Constrained Event Linking and Automatic Labeling for Cross-Document Accident Reports

Wenhua Zeng [1,2,*], Wenhu Tang [1,*], Diping Yuan [3], Bo Zhang [4] and Yuhui Zeng [2,*]

[1] School of Electric Power Engineering, South China University of Technology, Guangzhou 510641, China
[2] Shenzhen Urban Public Safety and Technology Institute, Shenzhen 518046, China
[3] Shenzhen Research Institute, China University of Mining and Technology, Shenzhen 518057, China; dpyuan2002@aliyun.com
[4] School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China; 4090@cumt.edu.cn
* Correspondence: zengwh@szsti.org (W.Z.); wenhutang@scut.edu.cn (W.T.); zengyh230025@stu.nfu.edu.cn (Y.Z.); Tel.: +86-755-66870324 (W.Z.)

Abstract

Constructing reusable accident-text corpora is hindered by anonymization, heterogeneous sources, and sparse labels, which complicate cross-document event linking. We propose a spatiotemporal lattice-constrained approach that encodes administrative hierarchies and temporal granularity, defines domain-informed consistency criteria, instantiates spatial/temporal relations via a subset of RCC-8 and Allen's interval algebra, estimates anchor weights via smoothing with monotonic projection, and fuses signals using a constrained monotonic network with explicit probability calibration. An active-learning decision rule—combining maximum probability with a probability-gap criterion—supports scalable automatic labeling, and controlled augmentation leverages instruction-tuned LLMs under lattice constraints. Experiments show competitive ranking (Hit@1 = 41.51%, Hit@5 = 77.33%) and discrimination (ROC-AUC = 87.34%), with the best F1 (62.46%). The method yields the lowest calibration errors (Brier = 0.14; ECE = 1.97%), maintains performance across sources, and exhibits the smallest F1 fluctuation across thresholds (Δ = 1.7%). In deployment-oriented analyses, it auto-labels 77.7% of cases with 97.51% accuracy among high-confidence outputs while routing 22.3% to review, where the true-positive rate is 81.46%. These findings indicate that integrating structured constraints with calibrated probabilistic fusion enables accurate, auditable, and scalable event linking for accident-corpus construction.

Keywords: event linking; accident reports; spatiotemporal lattice; automatic labeling; safety engineering; corpus construction

Check for updates

Received: 2 October 2025
Revised: 25 November 2025
Accepted: 31 December 2025
Published: 6 January 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Research on safety accidents has long been hindered by the lack of unified, standardized, and reusable public datasets. Constructing high-quality datasets is fundamental for advancing related research. However, manual collection and annotation of data are costly and time-consuming and do not scale to large datasets. Moreover, privacy-preserving regulations often remove or generalize key quasi-identifiers in accident reports, making it impossible to directly align structured accident records with narrative reports. This challenge falls within the domain of privacy-preserving record linkage (PPRL), which aims to link records while protecting sensitive information [1]. Consequently, accident-related

corpora face additional challenges compared to open-domain text tasks, including limited scale, diverse genres, and sparse annotations. Specifically, the challenges are as follows:

Diverse data sources and structures: Authoritative sources (e.g., accident statistics and direct reporting systems) provide accurate information but are limited in scope and volume. In contrast, semi-structured or unstructured texts (e.g., news reports, online content, internal documents) are abundant and rich in detail but exhibit wide variation in writing style, focus, and information granularity.

Cross-document event alignment difficulties: Structured records typically contain key attributes (event time, location, losses, responsible units), yet many public reports obscure or omit such identifiable cues due to privacy or compliance requirements, hindering alignment with structured data.

Long-tail label distribution: Accident categories follow a long-tail distribution, with a few common types dominating the dataset while many rare categories have few instances.

Therefore, effectively integrating heterogeneous multi-source data and leveraging structured information for scalable automatic annotation has become a core challenge in constructing accident analysis datasets. In the field of natural language processing (NLP), event linking aims to connect event mentions in text to corresponding event records in a knowledge base [2] and has long been a focal point in evaluation tasks such as ACE and TAC. Compared to entity coreference, event linking faces greater technical challenges, and the research is relatively underdeveloped. The core difficulties include cross-source narrative differences, lexical diversity, local information loss, and global consistency constraints.

The development of event linking has evolved from early rule-based methods [3–11] to traditional machine learning techniques [12–16] and, more recently, to deep learning methods [17–21]. With the rise of pre-trained large language models (LLMs), significant progress has been made in event linking methods based on reading comprehension [22], knowledge distillation [23], and instruction fine-tuning [24]. However, existing LLM-based approaches typically assume observable anchors and stable surface forms; faced with anonymized, fine-grained accident texts, they suffer from granularity mismatch (e.g., city vs. district), missing anchors (suppressed entities), and over-generalization—yielding confident yet poorly calibrated scores without structural guarantees. In contrast, a lattice-constrained formulation can recover interpretable alignment signals from residual spatiotemporal structure and maintain calibrated probabilities suitable for audit.

To address these challenges, this paper proposes an event linking method based on spatiotemporal containment consistency. The method uses structured records from accident statistics and reporting systems as anchors to accurately link event descriptions in open-domain texts. It models administrative hierarchies and temporal granularity in a unified spatiotemporal lattice, leveraging a subset of the Region Connection Calculus (RCC-8) [25] for spatial containment consistency and a subset of Allen's interval algebra [26] for temporal consistency. A probability fusion mechanism is introduced to integrate multiple confidence factors and output cross-document alignment probabilities with uncertainty measures, thereby effectively separating samples suitable for automatic annotation from those requiring human review. Additionally, based on these spatiotemporal constraints, we introduce a controlled data augmentation strategy using an instruction-tuned large language model (LLM) to automatically generate a large number of high-confidence labeled samples. Experimental results show that the proposed method can effectively align structured accident records with multi-source public reports, support scalable automatic annotation, and provide an expandable solution for efficiently constructing high-quality, auditable accident corpora. The main contributions of this study are summarized as follows:

1. A spatiotemporal containment consistency-based cross-document event linking method. By incorporating domain knowledge, we define a set of spatiotemporal

and type consistency criteria, compute anchor weights using smoothing and monotonic projection, and develop a multi-feature fusion scoring model to calculate the alignment probability between structured records and event mentions in online texts.
2. A sample selection strategy combining maximum alignment probability and probability gap. Under an active-learning framework, we introduce a selection criterion that considers both the maximum alignment probability and the probability gap, enabling high-confidence automatic annotation while maintaining high recall. This provides quantitative support for human–machine collaborative annotation.
3. The proposed method achieves performance comparable to the strongest baseline models in accident dataset construction tasks. It demonstrates its effectiveness and practicality across several key metrics.

The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 describes the proposed model and methods in detail; Section 4 introduces experimental design and datasets; Section 5 analyzes experimental results; Section 6 discusses the results; and Section 7 discusses the limitations of the method and future directions and concludes the paper.

2. Literature Review

Event linking is a core task that aims to associate event mentions in text with corresponding event nodes in a knowledge base [2]. The key challenges lie in handling cross-source narrative differences, lexical diversity, local information loss, and global consistency constraints. Research in this field has generally progressed through several stages: rule-based and template-driven methods, graph models with global inference, supervised learning with neural representations, and end-to-end cross-document modeling methods driven by pre-trained language models (PLMs). However, applying these PLM-driven methods to the specialized context of accident reports—which often involve anonymized fields, hierarchical spatiotemporal granularity, and domain-specific terminology—introduces additional challenges that general PLM approaches may not fully address.

Early research mainly relied on rule-based and heuristic matching methods. For example, Humphreys et al. [3] proposed a cross-document event alignment method based on template slots and manually defined rules, using "trigger word consistency + key argument overlap" as criteria for matching and suppressing erroneous matches through domain knowledge. Ahn [4] emphasized, from a task decomposition viewpoint, that the underlying "trigger word–argument" structure plays a crucial role in event coreference and linking. He highlighted that errors in initial event recognition can propagate through the pipeline and negatively affect downstream linking performance. Raghunathan et al. [5] introduced a multi-round filtering algorithm in entity coreference, in which rules are applied progressively from strict to broader criteria to merge mentions. Lu and Ng [6] later adapted this concept to event linking, significantly improving performance under limited data conditions. Cybulska and Vossen [7–9] proposed the "event bag" representation method, encoding events as sets of attribute slots and introducing a weighted slot matching mechanism that balances interpretability and robustness. Hovy et al. [10,11] further explored the theoretical boundaries of event identity and proposed a three-tier classification system ("same, different, quasi-same"), providing a theoretical foundation for subsequent studies on coreference and hierarchical event relationships. Although rule-based approaches can achieve high precision and are highly interpretable, they tend to generalize poorly across different domains and struggle to capture implicit semantic relationships.

Graph-based approaches represent event mentions as nodes in a graph, with edges encoding semantic similarity between mentions. They achieve globally consistent clustering by using algorithms such as spectral clustering, minimum-cut partitioning, or graph propa-

gation. Chen and Ji [12] proposed a two-stage framework ("feature combination → graph construction → graph partitioning") and clearly defined the process for cross-document event linking as "local scoring followed by global partitioning." In subsequent work [13], they applied a minimum-cut criterion to explicitly partition event clusters, thereby alleviating inconsistencies between local scoring and global clustering. Lee et al. [14] constructed a bipartite graph structure for entities and events, performing joint reasoning through shared arguments and optimizing clustering quality with an incremental merging strategy, significantly improving the accuracy and interpretability of cross-document alignment. Choubey and Huang [15] proposed an iterative expansion strategy guided by argument consistency. In each iteration, high-confidence event pairs are merged and similarity scores are updated, significantly improving recall—especially in multi-source news scenarios. Yang et al. [16] proposed a hierarchical generative model using non-parametric Bayesian methods. By incorporating prior constraints, this model guides the clustering structure and avoids the need to predefine the number of clusters. Graph-based methods provide strong global consistency and efficient information propagation. However, their performance heavily depends on the accuracy of the initial similarity estimates, and they often face high computational complexity when applied to large datasets.

With the introduction of annotated corpora such as ECB+ [27] and KBP [28], data-driven approaches have gradually become mainstream. Supervised learning methods typically treat event linking as a mention classification or clustering task and integrate features such as trigger words, semantic classes, and syntactic structures. Chen et al. [29] proposed a multi-parser ensemble framework to enhance the robustness of mention determination. However, these corpora cover open-domain or newswire texts, which may differ significantly from the specialized accident reports considered here, potentially affecting the direct applicability of methods. Lu et al. [30,31] systematically explored joint reasoning and multi-task learning approaches. They used Markov logic networks and multi-task learning to improve global consistency and introduced a probabilistic ranking mechanism to mitigate global inconsistencies. Bejan and Harabagiu [32,33] developed an unsupervised event coreference model using non-parametric Bayesian methods. Leveraging external semantic resources such as FrameNet, their model achieved performance comparable to supervised methods on open-domain texts. Araki and Mitamura [34] jointly trained trigger identification and event coreference using a structured perceptron, significantly reducing error propagation. Upadhyay et al. [35] critically examined cross-document evaluation metrics and advocated using multiple evaluation measures to provide a comprehensive assessment of model performance. Together, supervised and generative models have improved feature fusion and global reasoning. Techniques such as joint learning and ranking optimization have been key to transitioning from local mention matching to coherent global clustering.

Neural network methods significantly reduce the dependence on manual feature engineering by learning representations end-to-end. For example, Kenyon-Dean et al. [17] employed clustering-based regularization to train a BiLSTM encoder, drawing similar events closer in the embedding space and thus enhancing model robustness. Peng et al. [18] addressed event detection and coreference jointly under weak supervision, offering a feasible solution to the high annotation costs of event tasks. Barhom et al. [19] extended the span-based representation and cascade clustering mechanism from entity coreference to event linking. This approach achieved state-of-the-art performance on the ECB+ dataset while simplifying the traditional pipeline. Cremisini and Finlayson [20] conducted a systematic analysis and found that careful negative sampling and threshold tuning could achieve strong baseline performance, while some complex model components contributed minimally. Allaway et al. [21] proposed a sequential cross-document coreference method,

reducing inference complexity through state memory and historical consistency constraints, making it suitable for large-scale scenarios. The main advantage of neural methods lies in their ability to learn transferable semantic representations and support end-to-end optimization. Incorporating sequential reasoning further improves computational efficiency.

The emergence of pre-trained language models such as BERT has significantly enhanced semantic alignment and context modeling capabilities. For instance, Joshi et al. [36] applied BERT to entity coreference and demonstrated the model's ability to capture cross-sentence context. They later improved phrase-level representations in event modeling by using span masking strategies [37]. In cross-document coreference resolution tasks, Cattan et al. [38] proposed a transformer-based framework for linking events and entities across long documents. Yu et al. [39] utilized a shared encoder to jointly learn representations for events and their arguments, highlighting the importance of capturing document-level structural information in event linking. Zeng et al. [40] introduced paraphrasing resources and argument-aware embeddings to alleviate matching issues caused by lexical diversity. Caciularu et al. [41] proposed a Cross-Document Language Model (CDLM) that significantly improved consistency via multi-document masking pre-training. Longformer [42] and other long-text models have also been widely used for document-level encoding. Taken together, these key advances in the pre-trained language model era, including cross-document pre-training, span-level semantic enhancement, and long-context encoding, have significantly improved handling of lexical diversity and semantic equivalence.

In recent years, large language models (LLMs) have introduced new approaches to event linking. Some studies [43,44] evaluated the performance of GPT-series models in zero-shot and few-shot information extraction tasks, finding that while LLMs perform well in simple scenarios, they still significantly lag behind specialized models on more complex tasks. To combine the broad capabilities of LLMs with the strengths of smaller models, several works have attempted to integrate LLMs' semantic summarization ability with the task-specific fine-tuning of small language models (SLMs). For example, Min et al. [45] used GPT-4 to generate event summaries, then employed an SLM for coreference resolution. Nath et al. [23] generated reasoning explanations with an LLM and distilled these into a smaller model to improve interpretability and performance. Wang et al. [24] used instruction fine-tuning to unify multiple extraction tasks, while Ding et al. [46] enhanced the model's sensitivity to semantic differences by utilizing counterfactual samples.

However, despite their strong performance on several benchmarks, LLM-based methods still have limitations. Research [22,44,45] indicates that prompt-based methods struggle to cover complex annotation standards, and LLMs still show lower accuracy in event coreference compared to supervised models. Bugert et al. [47] found that feature engineering systems perform more robustly in cross-corpus testing, whereas purely neural models have poor generalization across domains. Together, these results suggest that effective annotation systems should incorporate corpus-specific features and appropriate constraint mechanisms to achieve better cost-effectiveness in annotation tasks.

3. Methodology

To address the issues of weakened explicit cues in publicly available accident texts due to anonymization, as well as significant genre and structural differences in cross-source texts, we frame our problem as a privacy-preserving record linkage (PPRL) task. Because anonymization often removes or generalizes key quasi-identifiers such as time and location, standard record linkage methods cannot directly align structured accident records with narrative reports. We therefore propose an event linking method based on spatiotemporal containment consistency to link accident descriptions across documents while preserving privacy. Structured records from accident statistics and direct reporting systems serve as

anchors for accurate matching with event description fragments in open-domain texts. The overall process, as illustrated in Figure 1, includes four core modules: feature extraction, consistency calculation, probability fusion, and decision output.

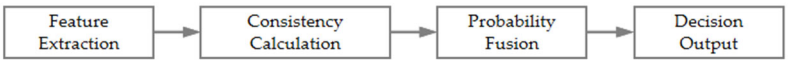

Figure 1. Overall Process. The process begins with feature extraction from structured records and narrative texts, followed by spatiotemporal lattice-based consistency calculation. A monotonic network then fuses multi-feature confidences into calibrated alignment probabilities, which support the final automatic labeling or human review decision via an active-learning rule.

First, we model administrative hierarchies and temporal granularity as a spatiotemporal lattice based on a subset of the Region Connection Calculus (RCC-8) [25] and Allen's interval algebra [26]. In this lattice, spatial containment relations (e.g., province–city–district–town) and temporal interval relations (e.g., contains, overlaps, before/after) are formalized as primitive operations. A set of spatiotemporal and type consistency criteria is defined based on domain knowledge. Next, anchor weights for each criterion are calculated through smoothing estimation and monotonic projection. Then, a multi-feature fusion scoring model is constructed, which integrates multi-factor confidence and outputs cross-document alignment probabilities with uncertainty measures. Finally, an active-learning framework combines a maximum-probability and probability-gap sample selection strategy, effectively delineating the boundary between automatically labeled samples and those requiring human review.

Additionally, to alleviate data sparsity, we propose a controlled data augmentation strategy leveraging an instruction-tuned large language model (LLM). Augmented samples are generated under strict spatiotemporal constraints to preserve privacy and semantic fidelity; each generated example is re-evaluated by the alignment model, and only those whose consistency scores remain within a predefined tolerance are retained. We also discuss potential risks of LLM-based augmentation, including semantic drift, fabrication, and bias, and incorporate manual or semi-automatic filters to mitigate these issues. The subsequent sections of this chapter will formalize the definition of each component of this method and provide a detailed explanation.

3.1. Symbol and Object Definitions

Let the set of accident records be $\mathcal{R} = \{r_j\}$. Due to privacy regulations, these structured records often have sensitive quasi-identifiers removed or generalized, so missing or aggregated values must be handled explicitly. Each record r_j represents a potential anchor event from the structured accident database and comprises three types of elements:

(1) Spatiotemporal elements $\left(r_S^j, r_T^j\right)$. The spatial location $r_S^j = (s_1, s_2, s_3, s_4)$ records the administrative hierarchy—province, city, district (or county), and street/town. The temporal attribute $r_T^j = [t_{\text{start}}, t_{\text{end}}]$ is a time interval indicating the event's occurrence window.

(2) Enumerative non-spatiotemporal elements r_E^j, such as accident severity level, accident type, and work scenario. Each field f_k corresponds to a hierarchical tree H_k defined by industry standards; for example, accident types may be arranged as "fire > electric fire > transformer fire".

(3) Structural elements r_X^j, including numeric attributes such as casualties or economic losses. These are stored in a vector $\mathbf{x}_j$; if a component is missing due to anonymization, a mask is applied to prevent penalizing the absence of information.

Let the set of accident texts from the internet and industry channels be $\mathcal{C} = \{c_i\}$, where each text c_i consists of a narrative description and additional scattered context.

Because these narratives are typically anonymized, they may omit specific times, locations, or names; therefore, we only extract candidate values and their confidences, rather than assuming a single true value. Using rule-based patterns and regular expressions, we derive multiple candidate sets from each text:

- A spatial candidate set $S_i = \left\{ \mathbf{c}_S^{i,k} \right\}$, where each candidate encodes up to four administrative levels, and missing levels are denoted by null symbols.
- A temporal candidate interval set $\mathcal{T}_i = \left\{ [a,b]^{i,\ell} \right\}$, representing possible time intervals extracted from the narrative.
- An enumerative field candidate set $\mathcal{E}_i = \left\{ c_{E,k}^{i} \right\}$, corresponding to possible values for fields like accident type or severity.
- A structured vector $\mathbf{x}_i$, empty if no numeric information is mentioned.

Each candidate is associated with an extraction confidence score: $q_S^{i,k}$, $q_T^{i,\ell}$ and $q_E^{i,k}$ quantify the reliability of the spatial, temporal, and enumerative candidates, respectively, as determined by the extraction algorithm.

The objective of cross-document accident linkage is to match each text c_i to its most likely structured record r_j. To this end, we compute an alignment probability $p_{i,j} = \Pr(r_j \mid c_i)$ for every pair of text and candidate record. A dual-threshold decision rule is then applied: if the maximum probability exceeds a confidence threshold η and the gap between the top two probabilities exceeds a separation threshold δ, we automatically label c_i as linked to r_j; otherwise, the case is flagged for human review. Both thresholds η and δ are dimensionless constants optimized on a validation set to balance false positives against annotation cost.

3.2. Spatiotemporal Lattice Structure and Consistency Criterion Modeling

One of our key innovations is to cast the matching problem in a formal spatiotemporal lattice that captures both spatial and temporal containment relationships. This lattice provides a common framework for evaluating the consistency of a candidate event with an anchor record across multiple dimensions. Specifically, we employ a subset of the Region Connection Calculus (RCC-8) to represent topological relations among administrative regions and a subset of Allen's interval algebra to represent temporal interval relations. These formal systems allow us to quantify how closely a candidate's location or time relates to the record's location or time (e.g., exact match, containment, overlap, disjointness). We collectively refer to these criteria as spatiotemporal containment consistency.

Spatial consistency quantifies how well a candidate location $\mathbf{c}_S^{i,k}$ matches a record location $\mathbf{r}_S^{j}$ on the administrative lattice. We classify the relationship between candidate and record into four categories (a subset of the RCC-8 relations):

– EXACT: all four administrative levels match (province, city, district, town);
– ANCESTOR: the candidate is a higher-level ancestor of the record (e.g., same city but no district specified);
– COUSIN: the candidate and record share a higher-level ancestor but differ at a lower level (e.g., same city, different districts);
– DISJOINT: no common ancestor exists.

Figure 2 visualizes these four relations.

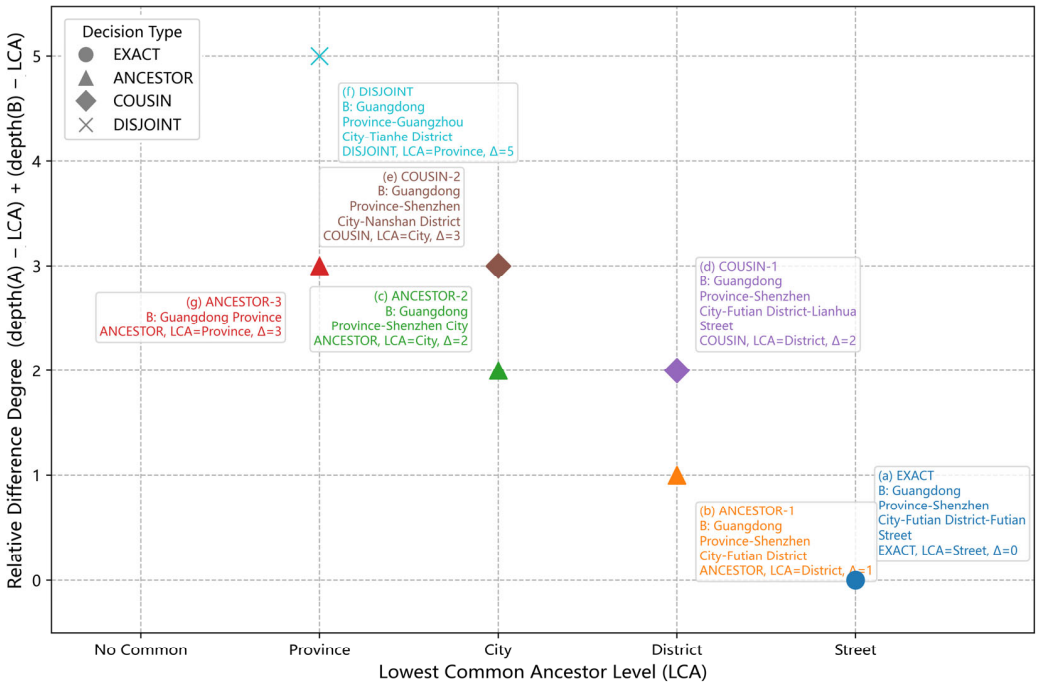

Figure 2. Spatially Constrained Judgment Matrix. It provides a unified, interpretable 2-D layout for auditing spatial decisions, calibrating weights, and diagnosing mismatch patterns. This diagram uses the reference location B set to "Guangdong Province–Shenzhen City–Futian District–Futian Street" and plots seven text-extracted candidate addresses A under the administrative-hierarchy matching scheme. The x-axis is the Lowest Common Ancestor (LCA) level with categories {No Common, Province, City, District, Street}. The y-axis is the relative difference degree, defined as $(\text{depth}(A) - \text{LCA}) + (\text{depth}(B) - \text{LCA})$, which quantifies how far the two locations differ below their LCA. Decision types from the RCC-8 subset are encoded by marker shapes: EXACT (circle), ANCESTOR (triangle), COUSIN (diamond), and DISJOINT (cross). Panel labels (a)–(g) illustrate typical cases: (a) EXACT (all four levels match the reference); (b)–(g) varying degrees of ANCESTOR or COUSIN divergence, and DISJOINT when no common ancestor exists at the province/city level.

We assign a weight $\omega_S^{\text{EXACT}} \geq \omega_S^{\text{ANCESTOR}} \geq \omega_S^{\text{COUSIN}} \geq \omega_S^{\text{DISJOINT}} \geq 0$ to each relation to reflect its importance. The spatial compatibility for text c_i and record r_j is defined as:

$$\phi_S(r_j, c_i) = \sum_k \omega_S^{(\text{rel}(\mathbf{c}_S^{i,k}, \mathbf{r}_S^j))} q_S^{i,k} \tag{1}$$

where $\text{rel}(\cdot,\cdot)$ returns the relation class between the k-th spatial candidate and the record, and $q_S^{i,k}$ is the extraction confidence of that candidate. This formulation ensures that more compatible candidates contribute higher scores.

Temporal consistency measures how well a candidate time interval aligns with the record's time interval. Based on a subset of Allen's interval algebra, we consider five relations.

- EXACT: the candidate interval exactly matches the record interval;
- CONTAINS: the candidate interval fully contains the record interval;
- OVERLAP: the intervals overlap, but neither contains the other;
- AFTER (or BEFORE): the candidate starts after the record ends (or vice versa);
- DISJOINT: no temporal relationship exists.

Figure 3 illustrates the five temporal relations.

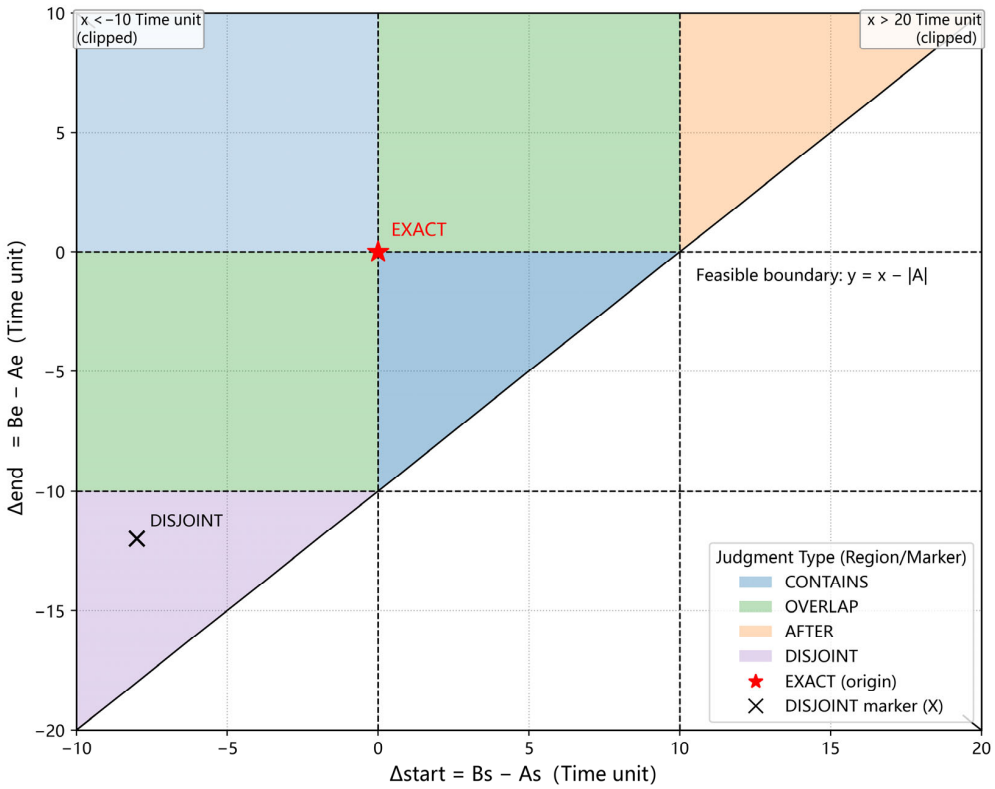

Figure 3. Temporally Constrained Judgment Matrix. It offers an intuitive visualization of temporal consistency ϕ_T, supports threshold selection, and facilitates inspection of contradictory or borderline interval configurations. This diagram places two time intervals on a common plane: candidate A $[A_s, A_e]$ and record B $[B_s, B_e]$. The x-axis is $\Delta_{start} = B_s - A_s$ and the y-axis is $\Delta_{end} = B_e - A_e$, measured in a consistent, explicitly stated time unit. The solid diagonal $y = x - |A|$ (with $|A| = A_e - A_s$) is the feasible boundary implied by interval length constraints. The plane is partitioned into the Allen-algebra subset of relations: EXACT at the origin (red star), CONTAINS (two wings), OVERLAP (banded regions), AFTER (upper-right), and DISJOINT (violet area outside the feasible band; an example is marked with "X").

We define weights $\omega_T^{\text{EXACT}}, \omega_T^{\text{CONTAINS}}, \omega_T^{\text{OVERLAP}}, \omega_T^{\text{AFTER}}, \omega_T^{\text{DISJOINT}}$ such that higher compatibility relations receive larger weights. If a record contains multiple possible time intervals (e.g., due to aggregated reports), we iterate $[a, b] \in T(r_j)$ over each interval to evaluate the relation and take the maximum compatibility. The temporal compatibility is then computed as:

$$\phi_T(r_j, c_i) = \max_{[a,b] \in T(r_j)} \sum_{\ell} \omega_T^{(\text{rel}([a,b],[a_\ell,b_\ell]^{i,\ell}))} q_T^{i,\ell} \tag{2}$$

where $q_T^{i,\ell}$ is the confidence of the ℓ-th temporal candidate.

Non-spatiotemporal consistency accounts for enumerative categorical fields, such as accident severity, type, or work scenario. Each field value corresponds to a node on a hierarchical tree H_k, where the tree structure captures categorical relations (e.g., "fire" > "electric fire" > "transformer fire"). Given a record value $r_{E,k}$ and a candidate value $c_{E,k}$, we calculate the distance on the tree $\text{dist}(r_{E,k}, c_{E,k})$, which is dimensionless and represents how closely the two values match. Non-spatiotemporal consistency is then defined using an exponential kernel function:

$$\phi_E(r_j, c_i) = \exp(-\lambda_k \cdot \text{dist}(r_{E,k}, c_{E,k})), \tag{3}$$

where λ_k controls the decay rate of similarity.

If the field is not mentioned in the text, we set $\phi_E = 1$, avoiding penalization for missing values, which ensures that the model does not incorrectly penalize the absence of information.

Structural information consistency measures the similarity of numeric attributes, such as casualties or economic losses. Let $\mathbf{x}_j$ be the record's numeric vector and $\mathbf{x}_i$ be the candidate's vector, with missing components masked by a binary vector $\mathbf{M}$. We compute the masked 1-norm difference:

$$d(\mathbf{x}_j, \mathbf{x}_i) = \sum_n M_n \cdot |x_{j,n} - x_{i,n}|, \tag{4}$$

which accounts for missing information using the mask vector $\mathbf{M}$.

This difference is then transformed into a compatibility score using an exponential decay function:

$$\phi_X(r_j, c_i) = \exp(-\varepsilon \cdot d(\mathbf{x}_j, \mathbf{x}_i)) \tag{5}$$

where $\varepsilon > 0$ controls the rate of decay.

If the candidate text does not provide numeric information (due to anonymization), we set $\mathbf{M} = \mathbf{0}$ and assign $\phi_X = 1$, avoiding penalization for missing data and allowing other features, such as time, location, and type, to dominate in matching. This approach ensures that missing or anonymized numeric values do not result in severe mismatches.

The spatiotemporal lattice structure and consistency criteria described above enable the system to handle noisy and incomplete data effectively. By combining spatial, temporal, and non-spatiotemporal consistency measures, the model produces stable and interpretable compatibility scores. This approach ensures that even when some information is missing or anonymized, the system can make reliable event linking decisions. These consistent frameworks, such as the RCC-8 and Allen subset, ensure that the model's decisions are robust and interpretable.

3.3. Weight Estimation and Monotonic Projection

To convert discrete relationships into comparable numerical weights, this study adopts the Laplace (Lidstone) smoothing method [48]. This method improves the stability of weight estimation, particularly when the sample size is small. Laplace smoothing is a regularization technique that adjusts empirical frequency estimates by introducing a prior, helping avoid extreme probability values of 0 or 1 due to insufficient samples. This technique has been widely used in tasks such as text classification, language modeling, and record linkage [49]. The core idea is to smooth empirical frequencies by introducing prior information, enhancing the robustness of the data fusion process [48].

Let $\mathcal{R}$ be a discrete relationship category, where the number of positive examples in the calibration set is denoted as N_{pos}, the number of negative examples as N_{neg}, and the total number of examples as N_{total}. If we directly use the empirical frequency $p_{\text{empirical}} = N_{\text{pos}}/N_{\text{total}}$, extreme estimates may arise, especially with small sample sizes. To mitigate this, we apply the Beta–Bernoulli model for smoothing. The prior α is set as $\alpha \in [0, 1]$, and the posterior expectation is:

$$p_{\text{smooth}} = (N_{\text{pos}} + \alpha)/(N_{\text{total}} + 2\alpha) \tag{6}$$

This smoothing formula ensures that the probability estimate remains bounded and less susceptible to overfitting or underfitting due to sparse data. The parameter α (a pseudo-count) controls the intensity of smoothing and can be adjusted for different levels of uncertainty.

Some relationships between features have an explicit order of strength. For example, temporal relationships generally follow the order "Exact match $\geq$ Upper-level

containment ≥ General overlap ≥ After official ≥ Disjoint," and spatial relationships usually satisfy "Exact match ≥ Upper-level containment ≥ Same ancestor, different branches ≥ No relation." Due to sampling fluctuations, independently smoothed weights may violate this natural order of relations. To address this issue, we apply a monotonic projection method to ensure that the weights respect the inherent ordering of relations. The projection process is formulated as:

$$\min_{\mathbf{w}} \sum_{i} (\mathbf{1}\{w_i > w_{i+1}\} \cdot \|w_i - w_{i+1}\|) \tag{7}$$

This minimization problem is efficiently solved using the Pool-Adjacent-Violators (PAV) algorithm [50,51], which adjusts the weights while maintaining the predefined order. After projection, the final weight vector **w** reflects the trend in the data and preserves the monotonicity of relationships.

It should be noted that the above method is not only applicable to temporal and spatial relationships but can also be used for other discrete mappings that require weighting. For example, relationships in enumerative non-spatiotemporal fields (e.g., accident level, accident type) can be smoothed using the Beta–Bernoulli model, with the order of categories adjusted through monotonic projection. This method ensures that categorical relationships like "same level > same ancestor > different branches" are handled in a statistically robust way, providing stable weights for the fusion model. For structured fields, the numeric compatibility has already been computed using the masked distance-exponential mapping, as discussed earlier, so these can be treated directly on the same scale as discrete weights.

The output of this section is a stable and comparable compatibility table for each discrete relationship. These compatibility scores will serve as inputs for the fusion model in Section 3.4, where they will be fused with other continuous compatibility values. By ensuring statistical robustness and monotonicity at this stage, we can achieve a fair trade-off between different features and produce reliable, interpretable alignment probabilities.

3.4. Multi-Feature Scoring Model

This section aims to fuse the compatibility values of different features on the same scale and output a calibrated alignment probability. Through the weight estimation and monotonic projection methods described in the previous section, various discrete relationships have been mapped to stable and dimensionless compatibility values. Based on these pre-processed features, the fusion module learns the interaction and trade-off mechanisms across features. The model structure is lightweight, highly interpretable, and maintains a monotonic non-decreasing property for each input component. Consistent with Sections 3.1 and 3.2, the input to the fusion module includes the spatial compatibility $\phi_S(r_j, c_i)$, the temporal compatibility $\phi_T(r_j, c_i)$, the set of enumerative compatibilities $\{\phi_{E,k}(r_j, c_i)\}$, the structural numeric compatibility $\phi_X(r_j, c_i)$, and the semantic compensation term $\phi_{\text{sem}}(r_j, c_i)$; all are normalized to $[0, 1]$ prior to fusion.

For any candidate pair (c_i, r_j), the input vector is defined as:

$$z_{i,j} = \left[\phi_S(r_j, c_i), \phi_T(r_j, c_i), \phi_{E,1}(r_j, c_i), \ldots, \phi_{E,K}(r_j, c_i), \phi_X(r_j, c_i), \phi_{\text{sem}}(r_j, c_i)\right]^\top \tag{8}$$

where ϕ_S and ϕ_T denote the spatial and temporal compatibilities; $\phi_{E,k}$ denotes the compatibility of the k-th enumerative field (with K fields); ϕ_X denotes the masked numeric (structural) compatibility; and ϕ_{sem} denotes a semantic compensation term derived from normalized text–record semantic similarity. By construction, every component lies in $[0, 1]$, so the vector dimensionality is $d = 2 + K + 2$.

The semantic compensation $\phi_{\text{sem}}(r_j, c_i) \in [0, 1]$ is computed from a normalized semantic similarity between the narrative of c_i and the description associated with r_j (e.g., cosine

similarity of sentence embeddings, linearly mapped to $[0, 1]$). It increases the score when two texts are semantically close, even if some quasi-identifiers are generalized or missing due to anonymization, but it cannot override hard spatiotemporal conflicts enforced by ϕ_S and ϕ_T. This provides an interpretable balance: lattice-based consistency yields transparent constraints, while ϕ_{sem} contributes robustness to wording variation.

To enhance numerical stability, a log-odds transformation is applied to each dimension:

$$\widetilde{\mathbf{z}}_{i,j} = \log \frac{\mathbf{z}_{i,j} + \epsilon}{\mathbf{1} - \mathbf{z}_{i,j} + \epsilon} \tag{9}$$

where $\epsilon > 0$ is a small constant.

This transformation expands values near $0/1$ in a controlled way, reduces saturation effects, and makes different compatibility components more comparable before they enter the fusion layer.

The fusion module uses a single hidden-layer feedforward architecture, ensuring non-negative weights and using non-decreasing activation functions to maintain monotonicity:

$$\begin{gathered} \mathbf{w}_1^{(+)} = \text{softplus}(\mathbf{w}_1),\ \mathbf{w}_2^{(+)} = \text{softplus}(\mathbf{w}_2). \\ s_{i,j} = \mathbf{w}_2^{(+)\top} \sigma_+ \left(\mathbf{w}_1^{(+)} \widetilde{z}_{i,j} + \mathbf{b}_1 \right) + b_2, \end{gathered} \tag{10}$$

where $\mathbf{w}_1$, $\mathbf{w}_2$, $\mathbf{b}_1$ and b_2 are trainable parameters. The weights are processed using the softplus function to ensure non-negativity.

Because all effective weights are non-negative and the activation is monotone, the output score is monotone in each input component. Hence, improving any single compatibility (e.g., making time or location more consistent) cannot decrease the overall score, which preserves interpretability.

The alignment probability is calibrated using a temperature τ scaling method:

$$p_{i,j} = \sigma \left(\frac{s_{i,j} + b}{\tau} \right) \tag{11}$$

where σ is the sigmoid, and b, τ are fitted on a validation set by minimizing log loss or Brier score.

We also monitor ECE to ensure that the predicted probabilities match empirical hit rates, which is critical for auditable decisions in PPRL. Calibration transforms the raw score into a probability with reliable uncertainty semantics.

The model is trained using a small-scale alignment labeled dataset, including human-verified samples and high-confidence pseudo-labeled samples. The loss function uses binary cross-entropy with an additional ℓ_2 regularization term:

$$\mathcal{L} = -\frac{1}{N} \sum_{(i,j)} \left[y_{i,j} \log p_{i,j} + (1 - y_{i,j}) \log (1 - p_{i,j}) \right] + \lambda \|\Theta\|_2^2 \tag{12}$$

where $y \in \{0, 1\}$ is the alignment label, and $\lambda > 0$ is the regularization coefficient. The hidden-layer width is set in a small range (e.g., 8–32) to control capacity; the optimizer is Adam with early stopping. All hyperparameters (including temperature) are selected on a validation set to avoid ad hoc choices.

To align with privacy-preserving record linkage, features exclude direct identifiers; only quasi-identifier-based compatibilities and semantic similarity (with lattice constraints) are used. After training, we assess feature contributions via local sensitivity or integrated-gradient-style attributions to support post hoc interpretability.

3.5. Inference, Decision-Making, and Uncertainty

At the inference stage, let the alignment probability between text c_i and candidate record r_j be $p_{i,j} \in [0,1]$. Define the Bag-Level Strength and Probability Gap as:

$$\mathbf{S_i} = \max_j p_{i,j},\ \mathbf{\Delta_i} = \mathbf{S_i} - \max_{j \neq \mathbf{j_i^\star}} p_{i,j},\ \mathbf{j_i^\star} = \operatorname{argmax}_j p_{i,j} \tag{13}$$

Here, $\mathbf{S_i}$ and $\mathbf{\Delta_i}$ (with $\mathbf{j_i^\star}$) are explicitly defined and computed from the calibrated $p_{i,j}$, ensuring the statistics retain probabilistic meaning for auditable PPRL decisions.

Intuitively, the strength summarizes the top alignment confidence for c_i, while the gap quantifies the separation between the best and the second-best candidates; both are computed from the calibrated probabilities in Section 3.4. Using calibrated $p_{i,j}$ ensures that these decision statistics retain probabilistic meaning, which is essential for auditable, privacy-preserving record linkage (PPRL).

The actual decision adopts a dual-threshold rule: when the strength exceeds η and the gap exceeds δ, automatically link c_i to the corresponding r_j; otherwise, mark it for human review. Both η and δ are dimensionless and tuned on a validation set to balance false positives and review cost; we also verify that decisions are well-calibrated by monitoring Brier score and expected calibration error (ECE). If the gap is small, manual review is required even when absolute strength is high.

For evidently conflicting candidate pairs, a one-time pre-calibration penalty is applied to prevent severely inconsistent matches from receiving high probabilities. We modify the raw fusion score as:

$$\mathbf{s'_{i,j}} = s_{i,j} - \gamma \cdot \mathbf{1}[(\text{temporal} \in \{\text{AFTER}, \text{DISJOINT}\}) \wedge (\text{spatial} = \text{DISJOINT})] \tag{14}$$

As Equation (14) indicates, if the temporal relationship is AFTER or DISJOINT and the spatial relationship is likewise DISJOINT, we subtract a fixed constant $\gamma > 0$ from the pre-calibration score and then calibrate the adjusted probability as outlined in Section 3.4.

The penalized probability is obtained via temperature scaling:

$$p_{i,j} = \sigma\left(\frac{\mathrm{s'_{i,j}} + b}{\tau}\right) \tag{15}$$

This "hard-conflict" guard leverages the RCC-8 and Allen subsets from Section 3.2 to enforce interpretable consistency when spatiotemporal evidence is contradictory.

When multiple texts point to the same record and satisfy the auto-alignment condition, merging avoids duplicate counting. We define the set of auto-aligned texts for record r_j as:

$$\mathcal{C}(\mathbf{r_j}) = \{c_i : \operatorname{argmax}_k p_{i,k} = j,\ \mathbf{S_i} \geq \eta,\ \mathbf{\Delta_i} \geq \delta\} \tag{16}$$

If $|\mathcal{C}(\mathbf{r_j})| > 1$, we sort by $\mathbf{S_i}$ in descending order, keep the top-confidence text as primary, and merge others as auxiliary evidence under the same record, preserving source and timestamps. Merging does not alter labels; it consolidates provenance for auditing.

To support long-term operation and audits, the inference pipeline records relation categories and smoothed weights; all compatibility scores; input vectors; fusion scores; final probabilities; η, δ; decisions; and merge identifiers. Logs use pseudonymized IDs and hashed tokens to fit PPRL constraints while retaining reproducibility. We continuously monitor Brier score and ECE; when calibration drift is observed, temperature parameters are re-estimated without retraining the combiner.

3.6. LLM-Based Data Augmentation

Let the set of automatically aligned samples in the current training slice be $\mathcal{A}_{\text{train}}$. From this pool we construct a stricter subset for augmentation:

$$\mathcal{A}_{\text{aug}} = \left\{ c_i \in \mathcal{A}_{\text{train}} : \mathbf{S_i} \geq \boldsymbol{\eta}_{\text{aug}}, \boldsymbol{\Delta_i} \geq \boldsymbol{\delta}_{\text{aug}} \right\} \tag{17}$$

where $\mathbf{S_i}$ and $\boldsymbol{\Delta_i}$ are defined in Section 3.5, and $\boldsymbol{\eta}_{\text{aug}}$, $\boldsymbol{\delta}_{\text{aug}}$ are stricter than the auto-labeling thresholds to ensure high-precision seeds.

Augmented texts are generated under domain constraints. We compile a domain knowledge base $\mathcal{K}$ (terminology, entity hierarchies) and use the enumerative ontologies H_k from Section 3.2. For each seed c_i aligned to r_j, we derive a constraint set $\mathcal{C}_{\text{lat}}(\mathbf{r_j})$ from the RCC-8 subset (spatial) and Allen subset (temporal), fixing the target topological and interval relations relative to r_j. This ensures that augmented texts preserve spatiotemporal consistency.

We employ an instruction-tuned generator $\mathcal{G}_{\boldsymbol{\theta}}$ with prompt templates $\boldsymbol{\pi_m}$ to produce diversified but consistent variants. For each seed c_i linked to r_j, we generate M candidates:

$$\widetilde{c}_i^{(m)} = \mathcal{G}_\theta\left(c_i \mid \boldsymbol{\pi}_m, \mathcal{K}, \mathcal{C}_{\text{lat}}\left(\mathbf{r}_j\right)\right), m = 1, \ldots, M \tag{18}$$

We apply three controlled transformations:

(i) same-class entity substitution (siblings under H_k);
(ii) parent–child rewriting (ontology level up/down) while keeping the RCC-8/Allen relations to r_j unchanged;
(iii) knowledge-guided paraphrases that avoid adding direct identifiers.

For each $\widetilde{c}_i^{(m)}$, we re-extract features and recompute ϕ_S, ϕ_T, $\{\phi_{E,k}\}$, ϕ_X, ϕ_{sem} and the calibrated probabilities $p_{\widetilde{i,j}}$ (pipeline as in Sections 3.4 and 3.5). A candidate $\widetilde{c}_i^{(m)}$ is accepted iff all of the following hold:

$$\begin{gathered} \operatorname{argmax}_j p_{\widetilde{i,j}} = j_i^\star,\ \left|\phi_S\left(r_j, \widetilde{c}_i^{(m)}\right) - \phi_S\left(r_j, c_i\right)\right| \leq \kappa_S, \left|\phi_T\left(r_j, \widetilde{c}_i^{(m)}\right) - \phi_T\left(r_j, c_i\right)\right| \leq \kappa_{\mathbf{T}} \\ \sum_k \left|\phi_{E,k}\left(r_j, \widetilde{c}_i^{(m)}\right) - \phi_{E,k}\left(r_j, c_i\right)\right| \leq \kappa_E,\ p_{\widetilde{i,j_i^\star}} \geq \rho_{\text{aug}} \end{gathered} \tag{19}$$

where κ_S, $\kappa_{\mathbf{T}}$, κ_E are per-dimension tolerances (no degradation beyond a small margin) and ρ_{aug} is a probability floor for retention.

We enforce PPRL-compatible safeguards during augmentation: blocklisting direct identifiers (names, precise addresses), no finer-grained geocoding than the anchor's administrative level, time expressions preserved/aggregated to keep the Allen relation, toxicity/bias filters, and manual spot checks on a fixed budget. Rejected candidates are discarded and not logged beyond aggregate statistics.

Accepted augmented texts are added only to the training set with a sample weight $\mathbf{w}_{\text{aug}}$ (smaller than original samples to reduce confirmation bias); validation and test sets remain untouched. All provenance is logged: seed ID, r_j, template, constraint set, pre/post compatibilities, and acceptance decisions, supporting audits and ablations.

This constrained augmentation increases coverage while preserving the spatiotemporal relations to r_j and the calibrated alignment behavior of the model. By combining lattice-aware prompts, per-component tolerance checks, and a probability floor, we mitigate semantic drift and privacy risks while yielding diverse, label-consistent training texts.

3.7. Model Training and Implementation Details

The training process is divided into two stages. In the first stage, a small number of manually verified samples together with high-confidence samples obtained through strict rule-based filtering are used to lightly train the monotonic combiner described in Section 3.4, producing the initial probability outputs. Combined with the dual-threshold rule of Section 3.5, this completes the first round of large-scale automatic annotation. This stage generates a high-precision auto-labeled pool for subsequent augmentation and retraining while keeping validation/testing untouched.

Based on the accident occurrence time, the data are then split into training, validation, and test sets to avoid temporal leakage. We denote the chronological splits as $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{val}}$ and $\mathcal{D}_{\text{test}}$, respectively, where all hyperparameters are selected on $\mathcal{D}_{\text{val}}$ and final metrics are reported only on $\mathcal{D}_{\text{test}}$. The auto-labeled pool used for augmentation is limited to $\mathcal{D}_{\text{train}}$.

In Stage-1, we train the monotone combiner with non-negative weights and monotone activations (see Section 3.4) to obtain raw scores $s_{i,j}$ and calibrated probabilities $p_{i,j}$. We then apply the dual thresholds η and δ from Section 3.5 to each text c_i to produce auto-labels, yielding a seed pool $\mathcal{A}_{\text{train}} \subset \mathcal{D}_{\text{train}}$ and its stricter subset for augmentation $\mathcal{A}_{\text{aug}}$ as defined in Section 3.6.

In Stage-2, only the fusion module is fully trained and re-calibrated on $\mathcal{D}_{\text{train}} \cup \mathcal{A}_{\text{aug}}$; the validation and test sets are never augmented or re-estimated. Accepted augmented samples (see Section 3.6) are included with a weight $\mathbf{w}_{\text{aug}} < 1$ to mitigate confirmation bias; the temperature parameter τ is refit on $\mathcal{D}_{\text{val}}$ using Brier/ECE.

A unified preprocessing pipeline is applied:

(i) Text normalization (numbers/units, temporal expressions, place names);
(ii) Feature extraction (temporal anchors and granularity, administrative levels, enumerative field nodes, structural vectors, semantic similarity);
(iii) Consistency score computation and caching (columnar storage indexed by sample–candidate pairs).

Candidate retrieval uses an inverted index keyed by time windows, administrative hierarchy, and event type; the per-text candidate cap is K, tuned on $\mathcal{D}_{\text{val}}$ to balance recall and cost; and vectorized batch computations are used to accelerate scoring.

The first-stage training uses the Adam optimizer with a fixed learning rate and weight decay, batch size of 256, and early stopping with a patience of five epochs. The model input consists of the log-odds transforms of compatibility scores (see Section 3.4). A single-hidden-layer monotonic network with softplus activations ensures non-negative effective weights (see Section 3.4). Hyperparameters are selected on $\mathcal{D}_{\text{val}}$ via grid or line search; final settings are fixed before $\mathcal{D}_{\text{test}}$ evaluation.

With an average of K candidates per text and N texts, candidate retrieval/scoring is $O(NK)$; the fusion layer training is approximately $O(E \cdot d)$ with E training pairs and input dimension d. Static fields and intermediate compatibility are cached to avoid re-computation; matrix-batch operations are used end-to-end to reduce wall-clock time and memory traffic.

Experiments are conducted in Python 3.10 and PyTorch 2.x with fixed random seeds; both data and models are version-controlled. For auditability, we log relation categories and smoothed weights, all ϕ_* scores, input vectors $\mathbf{z}_{i,j}$, fusion scores $s_{i,j}$, probabilities $p_{i,j}$, thresholds η, δ, decisions, and merge IDs (Section 3.5). If calibration drift is detected (Brier/ECE), τ is re-estimated on $\mathcal{D}_{\text{val}}$ without retraining.

Final evaluation on $\mathcal{D}_{\text{test}}$ includes accuracy-type measures and calibration (Brier/ECE). We also conduct ablations, disabling one component at a time (lattice constraints, smoothing/projection, temperature calibration, semantic compensation) to quantify each contribution.

4. Experimental Setup

4.1. Data Sources and Processing

The experiments rely on two complementary sources. The first is a collection of structured accident records exported from authoritative statistics and direct reporting systems, denoted by $\mathcal{R}$. The second is a multi-source narrative corpus gathered from public web and industry channels, denoted by $\mathcal{C}$. To align with privacy-preserving record linkage (PPRL), all direct identifiers are removed or generalized before access; only quasi-identifiers (time, location, type) are retained at coarsened granularity.

Each record $r_j \in \mathcal{R}$ comprises the spatial vector $r_S^j = (s_1, s_2, s_3, s_4)$ over administrative levels (province, city, district/county, street/town); the temporal interval $r_T^j = [t_{\text{start}}, t_{\text{end}}]$; enumerative non-spatiotemporal fields $r_E^j = \{f_k\}$ mapped to domain ontologies H_k; and structural numeric attributes $r_X^j = \mathbf{x}_j$ (e.g., casualties, losses) accompanied by a mask for missing components. All schemas follow public safety standards; enumerative fields are normalized to node IDs in H_k to enable tree-distance computations introduced in Section 3.2.

For each text $c_i \in \mathcal{C}$, we apply normalization (numbers/units, sentence segmentation), then extract candidate sets using rules and regular expressions: the spatial set $S_i = \left\{\mathbf{c}_S^{i,k}\right\}$ (up to four administrative levels, missing levels left null), the temporal set $\mathcal{T}_i = \left\{[a,b]^{i,\ell}\right\}$, and the enumerative set $\mathcal{E}_i = \left\{c_{E,k}^i\right\}$, plus the structural vector $\mathbf{x}_i$ if present. Each candidate carries an extraction confidence $q_S^{i,k}, q_T^{i,\ell}, q_E^{i,k}$. We perform web-scale de-duplication and near-duplicate detection (shingling + locality-sensitive hashing) to suppress repeated reports while preserving provenance. Direct identifiers, if any, are masked prior to storage to comply with PPRL.

Location strings are canonicalized to the administrative hierarchy and geocoded to level IDs. In preprocessing, we also derive topological relations using a subset of RCC-8 (e.g., EXACT/ANCESTOR/COUSIN/DISJOINT) between $\mathbf{c}_S^{i,k}$ and r_S^j, so that spatial compatibility ϕ_S can be computed consistently at inference. Temporal expressions are normalized to closed intervals. We compute interval relations using a subset of Allen's algebra (EXACT/CONTAINS/OVERLAP/BEFORE/AFTER) between $[a,b]^{i,\ell}$ and r_T^j, enabling temporal compatibility ϕ_T. These lattice-aware features operationalize the theoretical framework in Section 3.2 during data processing.

Enumerative fields extracted from c_i are mapped into H_k via dictionaries and alias tables, with synonym consolidation and out-of-vocabulary back-off to parent nodes to avoid brittle mismatches. Structural attributes (if mentioned) are parsed into $\mathbf{x}_i$ with unit normalization; a binary mask is retained to indicate missing components so that ϕ_X in Section 3.2 applies no penalty when data are anonymized or absent.

We sort all items by occurrence time and partition them into $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{val}}$ and $\mathcal{D}_{\text{test}}$ to avoid temporal leakage. Only $\mathcal{D}_{\text{train}}$ is used for auto-labeling and LLM-based augmentation; $\mathcal{D}_{\text{val}}$ is used for hyper-parameter/temperature tuning; $\mathcal{D}_{\text{test}}$ is strictly held out. We have collected 1328 structured accident records and 6865 narratives collected between 2014 and 2024, across 8 accident types. Each record includes a 4-level spatial hierarchy vector, a time interval, enumerative fields, and numeric attributes. All baselines are processed with the same normalization, candidate extraction, and annotation pipeline (symbols/objects, consistency scores, dual-threshold decision) and use the same candidate cap K to match recall. Scripts are version-controlled; logs store only pseudonymized IDs and hashed tokens to comply with PPRL while preserving reproducibility.

4.2. Baseline Models

We select baseline models under a small-sample, two-stage training and calibration setting. All comparison methods share the same data splits, candidate generation pipeline, and evaluation protocol as the proposed method. On the input side, accident summary texts are uniformly used as the alignment target, denoted as c_i, while direct-report records are converted into brief descriptions by concatenating structured elements (time, location, category, casualties, etc.) as the alignment reference, denoted as r_j. On the output side, ranking quality, classification quality, and probability calibration quality are reported uniformly, using the calibrated alignment probability $p_{i,j}$ from Section 3.4, and thresholds η and gap δ as well as temperature-scaling parameters (b, τ) are determined only on the validation set.

To ensure reproducibility and fairness, the baselines adopt the same processing and annotation pipeline as the proposed method (symbol and object definitions, consistency score computation, dual-threshold decision rules, etc.) and maintain the same candidate pool size and recall target. All methods apply the same dual-threshold decision on $p_{i,j}$: auto-label if $p_{i,j} \geq \eta$ and the top-two probability gap $\geq \delta$; otherwise, send to review (cf. Section 3.1). Probability calibration quality is reported with Brier score and ECE, consistent with Section 3.4. This unified setup guarantees comparability under the same task formulation, reducing biases introduced by differences in data handling or engineering components.

The rule–probability baseline implements event linking under the Fellegi–Sunter (F-S) probabilistic record linkage framework [52]. It maps each summary–record sample to a field-level comparison vector, separately evaluating time interval relations, administrative-level containment relations, hierarchical distances of enumerative non-spatiotemporal elements, and consistency of structured vectors, then computing discrete or continuous evidence components. Based on a small calibration set, it estimates the conditional distributions of true matches and non-matches to generate likelihood ratios and produces "match," "non-match," or "undecided" judgments via upper and lower thresholds. For comparability, we convert the F-S scores to calibrated $p_{i,j}$ via temperature scaling (b, τ) on the validation set and apply the same η, δ decision as above. This method focuses on element-level evidence, offering an interpretable performance upper bound suitable for evaluating the limit when only structured consistency is used. The implementation retains interpretability for each component and jointly tunes thresholds and rejection strategies on the validation set to remain consistent with the probability calibration and risk-control standards of the proposed method.

The retrieval–re-ranking baseline uses the Cross-Document Event Coreference Search (CDECS) method [53], which transforms cross-document event coreference into a retrieval–ranking task and supports zero-shot or small-sample processes. Candidate recall is performed by an unfine-tuned general semantic encoder, followed by lightweight threshold adjustment and temperature scaling on the validation set for calibration. In this task, the accident summary is encoded as a query vector, and the textualized description of direct-report records is encoded as a library vector; nearest-neighbor recall is first performed, then similarity scores or a lightweight classification head are used for ranking and probabilistic scoring. We report probabilities as $p_{i,j}$ after temperature scaling (b, τ), and apply the same η, δ rule; initial recall uses top-K_{ret} nearest neighbors.

The zero-shot linking baseline employs the dual-tower dense retrieval model from the BLINK framework [54]. BLINK decomposes the "mention–entity linking" task into dual-tower retrieval and cross-encoder fine ranking, relying primarily on entity description texts rather than task-specific supervision, thus operating in zero-shot conditions. For this task, direct-report records are normalized into event-element descriptions (short texts of time, location, category, consequences, etc.) to build the vector index; accident summaries are encoded as query vectors for top-K_{ret} candidate retrieval and ranking, and similarity is mapped to probabilities $p_{i,j}$ via temperature scaling on the validation set. The same η, δ decision is used.

The generative zero-shot decision baseline adopts the Argument-Aware Event Linking (AAEL) method [23]. Without human annotation, it uses argument-aware prompts to structure the input so that the model focuses more on matching time, location, and participants during ranking. Applied to this task, summary–record pairs are filled into a prompt template, and the model outputs a judgment on whether they describe the same event and a brief rationale, which we convert to $p_{i,j}$ using the same (b, τ) calibration on the validation set, followed by the η, δ decision rule to align with the evaluation standard of this paper. This baseline tests the transferability and potential risks of large language models in cross-document event linking under low-supervision conditions.

The teacher–student distillation baseline uses an LLM-based method proposed at NAACL 2024 [55]. Its core idea is to have an autoregressive large language model generate free-text rationales as distant-supervision signals and distill them into a lightweight student model, thereby improving cross-document event coreference performance without increasing manual annotation. In this task, the teacher model generates alignment/non-alignment rationales and soft labels for summary–record samples, while the student model trains a classification head under small-sample conditions using a joint loss (hard-label cross-entropy plus soft-label KL divergence [56]). During inference, only the student model is used, with temperature scaling on the validation set. The output probabilities are reported as $p_{i,j}$ and thresholded with η, δ for consistency with our evaluation protocol. This baseline operates stably under few-label and pseudo-label settings and can test the feasibility and benefits of transferring teacher-model capabilities and achieving efficient student-model inference.

Together, these five baselines cover interpretable evidence-based probabilistic linking, retrieval–re-ranking under general semantic paradigms, dense retrieval zero-shot linking, argument-aware generative re-ranking, and "free-text rationale + distillation" low-annotation paradigms. All have peer-reviewed empirical results on public benchmarks (e.g., ECB+ [27]) and require only minimal adaptation when transferred to the one-to-one "accident summary–direct report" alignment task, making them suitable references under the small-sample and zero-shot conditions of this study. All metrics and decisions are computed on the calibrated $p_{i,j}$ with (b, τ) from Section 3.4 and the same η, δ rule.

4.3. Parameter Settings

Parameter settings include only points directly related to experiment reproducibility and adopt consistent selection principles across methods whenever possible. All methods share the same data splits, extractor versions, and candidate generation pipeline. Candidate pool size is determined on the validation set according to the criterion of "minimizing the average number of candidates while ensuring target recall" and is kept unchanged on the test set. All probability outputs are reported as the calibrated alignment probability $p_{i,j}$ and use temperature scaling with parameters (b, τ) fitted once on the validation set to minimize log loss or the Brier score; decisions follow the same dual-threshold rule with η (confidence) and δ (top-two gap) as in Sections 3.1–3.4; no adjustments are made during testing.

Text truncation strategies are kept consistent across methods: only "accident summary" and "record synopsis" are used, with priority given to preserving time, location, and key numerical expressions to avoid biases introduced by length differences.

For the Fellegi–Sunter baseline, key parameters include the prior smoothing strength of components and the likelihood ratio thresholds. Component conditional distributions are estimated from the small calibration set using symmetric Beta–Bernoulli smoothing, and the smoothing coefficient is selected on the validation set by grid search to jointly optimize the Brier score and ECE. For time and administrative-level relationships, component weights are subjected to isotonic projection using equidistant regression to preserve the order of strength, applied to the spatiotemporal relation weights $\omega_S^{(\cdot)}$ and $\omega_T^{(\cdot)}$ for consistency

with Sections 3.2 and 3.3. The likelihood ratio adopts joint upper and lower thresholds, with threshold selection primarily maximizing F1 while controlling the false positive rate within a preset range on the validation set. The log-likelihood ratio is linearly mapped to log-odds before temperature scaling and converted to $p_{i,j}$ with the same (b, τ) to ensure comparability of probability outputs across methods.

For the CDECS baseline, the core parameters are the encoder, similarity temperature, and recall depth. The encoder uses a publicly available general semantic model and remains frozen; similarity is fitted to a single temperature parameter on the validation set to map cosine similarity stably to log-odds; recall depth matches the candidate pool size, with no separate optimization per method to avoid selection bias. If small-sample fine-tuning is required, only a few contrastive learning steps are performed on positive–negative pairs from the calibration set, with learning rate and steps chosen on the validation set; if no fine-tuning is performed, the encoder remains frozen and only temperature and thresholds are fitted on the validation set. All reported probabilities are calibrated $p_{i,j}$ with (b, τ), and the η, δ decision is applied uniformly.

For the BLINK dense retrieval baseline, zero-shot operation is emphasized. The dual-tower encoder uses public weights and remains frozen; the entity side (here, the "record synopsis" side) builds the vector index, and the query side is the "accident summary." Under zero-shot settings, the cross-encoder re-ranking is disabled, relying only on dual-tower similarity ranking, with temperature scaling and threshold fitting on the validation set. Approximate retrieval parameters of the index are aligned with the candidate pool size, and no additional model-specific tuning is performed. If minimal adaptation is applied, it is limited to standardizing the textual format of "record synopsis" templates within the validation set without updating encoder parameters. Outputs are presented as $p_{i,j}$ after (b, τ) calibration, and decisions use the same η, δ rule.

For the LLM-based zero-shot re-ranking baseline, a deterministic inference configuration is used to reduce variance. The generation temperature is set to zero or near zero, sampling is disabled, and the maximum generation length is sufficient to cover the judgment and brief rationale. Inputs use a fixed pairwise template explicitly specifying time, location, and field elements; outputs are parsed into binary decision scores, then normalized to probabilities on the validation set using logistic regression or temperature scaling. To avoid prompt drift, the template remains unchanged throughout the experimental period; if refusals or "cannot judge" outputs occur, they are uniformly treated as the lowest-confidence negative examples, and this rule is consistently applied on the validation set without ad hoc modifications on the test set. For comparability, final probabilities are calibrated as $p_{i,j}$ via (b, τ), followed by the η, δ decision.

For the "rationale + distillation" baseline, a teacher–student two-stage configuration is adopted. The teacher is an open-source autoregressive large model used only to generate alignment/non-alignment rationales and soft labels offline and does not participate in online inference. The student is a long-context encoder with a lightweight classification head, trained only on the calibration set and high-confidence pseudo-labels under small-sample conditions, using a weighted sum of hard-label cross-entropy and soft-label KL divergence [56] as the loss function. The weights of the two losses, learning rate, training steps, and early-stopping criteria are all chosen by small grid search on the validation set, aiming to maximize F1 and Hit@k without sacrificing the Brier score or ECE. The distilled student runs independently at test time and uses the same temperature scaling and thresholds without invoking teacher inference. At evaluation, we report calibrated $p_{i,j}$ with (b, τ) and apply the shared η, δ rule.

To ensure fair comparison, the above settings follow two principles:

(i) any decision-related hyperparameters are chosen only on the validation set, with no temporary adjustments on the test set;
(ii) method-specific engineering optimizations are restricted within feasible bounds to avoid incomparability due to different candidate pool sizes, different text truncation strategies, or different external resources.

4.4. Evaluation Metrics

We adopt the following evaluation suite: Hit@k, ROC-AUC, Precision, Recall, F1, the Brier score, and Expected Calibration Error (ECE). All probability-based metrics are computed on the calibrated alignment probability $p_{i,j}$ (cf. Section 3.4). Ranking quality is assessed with Hit@k (Top-k hit rate) and ROC-AUC. Hit@k evaluates whether the ground-truth target appears within the top-k positions of the candidate list sorted by $p_{i,j}$; common choices of k are 1, 5, and 10, which reflect the model's best-match capability and the extent of candidate coverage useful during human review. Here k denotes the rank cut-off for Hit@k and should not be confused with the enumerative field count K in Section 3.4 or the retrieval depth K_{ret} used during candidate recall. ROC-AUC measures the robustness of the overall ranking, evaluated over $p_{i,j}$ for positive and negative pairs, with higher values indicating stronger discrimination between positive and negative instances. Precision, Recall, and F1 are computed from binary decisions obtained by the dual-threshold rule on $p_{i,j}$: predict "linked" if $p_{i,j} \geq \eta$ and the top-two probability gap $\geq \delta$ (parameters fixed from the validation set). The Brier score computes the mean squared error between predicted probabilities and the true labels, applied to $p_{i,j}$, where lower values indicate more accurate probability estimates; ECE is computed on $p_{i,j}$ using a fixed binning scheme shared across methods and assesses calibration by comparing predicted probabilities with empirical accuracies across confidence bins. These two metrics provide important guidance for setting decision thresholds between automated processing and human review in practical deployments and align with the temperature-scaling setup (b, τ) established on the validation set in Section 3.4.

5. Experimental Results and Analysis

This section reports and analyzes the main-task alignment under a few-shot, two-stage setting. To maintain evaluation consistency, all methods adopt the same data split, candidate-set generation pipeline, and input truncation strategy. Decision thresholds and temperature-based probability calibration parameters are fit on the validation set and then fixed on the test set to avoid leakage and ensure reproducibility.

5.1. Overall Performance Analysis

Table 1 summarizes the performance of each method on ranking metrics (Hit@1, Hit@5), classification metrics (Precision, Recall, F1), and probability metrics (ROC-AUC, Brier, ECE). For transparency and comparability, all metrics are computed under the same split and candidate pipeline as specified in Section 4. Our method achieves the highest F1 (62.46%) and the best calibration (Brier = 0.14, ECE = 1.97%), while maintaining competitive ranking (Hit@1 = 41.51%, Hit@5 = 77.33%); ROC-AUC is 87.34%, close to the best (87.56% by CDECS). These results indicate strong accuracy with reliable probabilities for automatic labeling.

From the overall comparison, the proposed method performs well on several key metrics, showing competitive results particularly on Hit@1, F1, ROC-AUC, and the calibration metrics (Brier and ECE). Concretely, our method obtains Hit@1 = 41.51% and Hit@5 = 77.33%, F1 = 62.46%, ROC-AUC = 87.34%, Brier = 0.14, and ECE = 1.97%, representing the best F1 and calibration among all systems while keeping discrimination comparable to the strongest baseline.

Table 1. Overall Performance Metrics (%).

Method	Hit@1	Hit@5	ROC-AUC	Precision	Recall	F1	Brier	ECE
F-S [52]	33.49	69.31	83.24	**74.83**	46.97	57.71	0.17	3.89
CDECS [53]	38.58	74.71	**87.56**	66.71	52.43	58.71	0.15	2.79
BLINK [54]	36.84	72.08	84.55	63.18	50.62	56.21	0.16	3.16
AAEL [23]	**42.12**	75.63	87.06	69.79	53.39	60.5	0.15	2.67
LLM [55]	40.93	**77.88**	86.69	67.12	**55.31**	60.65	0.15	2.52
Ours	41.51	77.33	87.34	73.92	54.07	**62.46**	**0.14**	**1.97**

In terms of ranking, AAEL achieves the best Hit@1 (42.12%), as shown in Figure 4a, followed by our method at 41.51% (a gap of 0.61%). The remaining methods rank as LLM (40.93%), CDECS (38.58%), BLINK (36.84%), and F-S (33.49%). For Hit@5, LLM leads with 77.88%, while our method attains 77.33% (a gap of 0.55%); AAEL (75.63%), CDECS (74.71%), BLINK (72.08%), and F-S (69.31%) follow. While Hit@k reflects top-rank retrieval quality, small deltas in ranking do not necessarily translate into more reliable automatic labeling, which depends critically on probability calibration and thresholding.

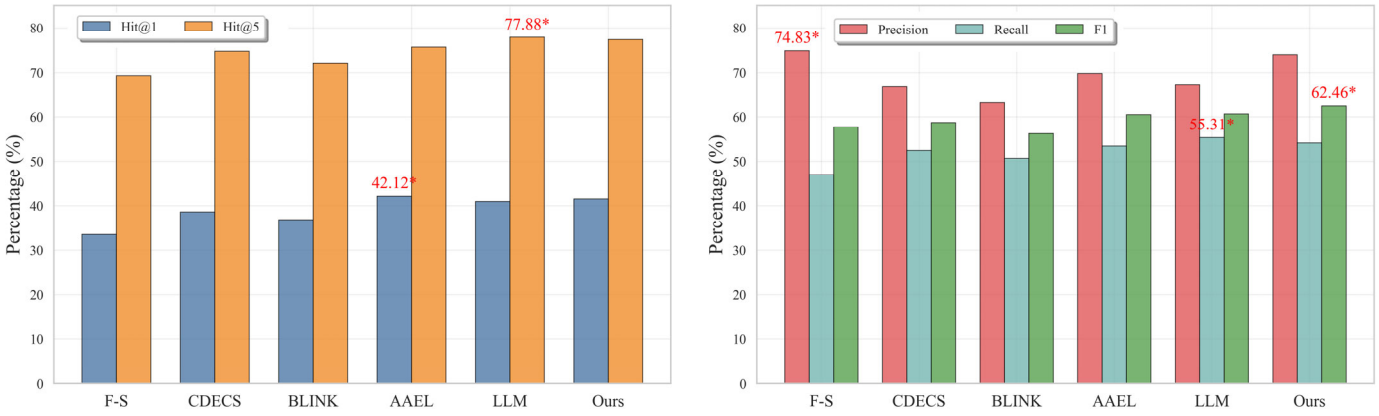

Figure 4. Performance comparison of baseline models and the proposed method. (**a**) Ranking performance (Hit@1 and Hit@5). (**b**) Classification performance (Precision, Recall, and F1). While some baselines achieve marginal leads in individual ranking metrics, the proposed method attains the best overall F1 score and a more favorable Precision–Recall balance, which is critical for reliable automatic labeling in practice. Numbers in red with asterisks (*) indicate the best performance in each metric.

Regarding classification metrics, as shown in Figure 4b, our method attains the highest F1 score (62.46%), exceeding the next-best LLM (60.65%) by 1.81%. Further analysis shows our Precision is 73.92%, second only to F-S (74.83%); our Recall is 54.07%, 1.24% lower than the best LLM (55.31%). Under a unified threshold across systems, this operating point offers a favorable Precision–Recall balance for auto-labeling, whereas F-S favors Precision at the expense of Recall and LLM does the opposite.

In probability calibration, as shown in Figure 5, our method achieves the lowest ECE (1.97%) and Brier score (0.14) among all methods. The Brier score measures the mean squared error of probabilistic predictions, where smaller values indicate predictions closer to the true distribution. ECE reflects calibration error; smaller values indicate more reliable confidence. Good calibration enables direct mapping from predicted probabilities to expected accuracy, thereby supporting threshold selection between automatic labeling and human review. For probability-based discrimination, ROC-AUC evaluates overall separability across thresholds; CDECS performs best (87.56%), and ours is a close second (87.34%, $\Delta = 0.22$ pp), so pairing strong discrimination with superior calibration is crucial for dependable PPRL-style deployment.

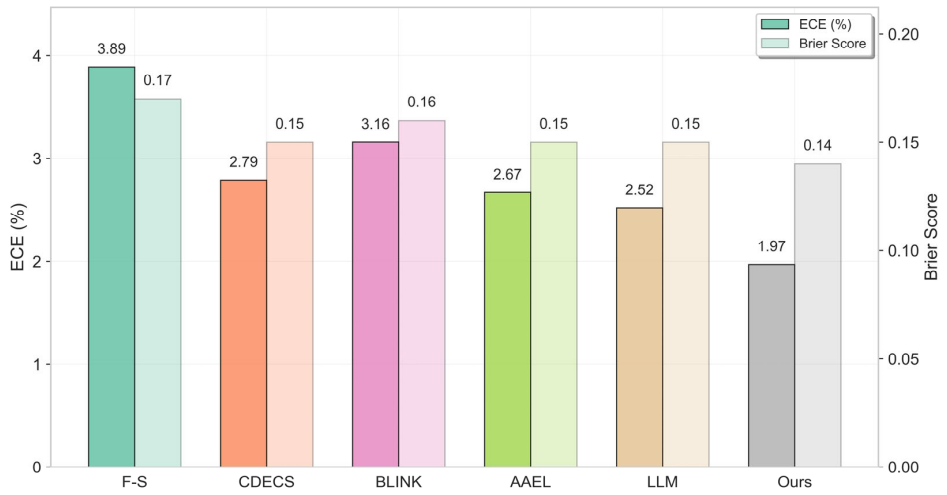

Figure 5. Comparison of probability calibration metrics (Brier score and Expected Calibration Error—ECE). The proposed method achieves the best calibration performance (lowest Brier score and ECE), indicating that its predicted alignment probabilities are the most reliable. This superior calibration is crucial for setting trustworthy decision thresholds between automatic labeling and human review in practical deployment.

Figure 6 depicts the probability calibration of each model. The x-axis denotes the mean predicted probability within a bin, and the y-axis the empirical positive rate. The ideal calibration curve (black dashed line) corresponds to perfect probability estimates. Results show clear differences across models: our method's curve lies closest to the diagonal, indicating high agreement between predicted and true probabilities; traditional methods (e.g., F-S) fall clearly below the diagonal, revealing systematic overestimation; some models (e.g., CDECS, BLINK) underestimate in low-probability bins but overestimate in high-probability bins.

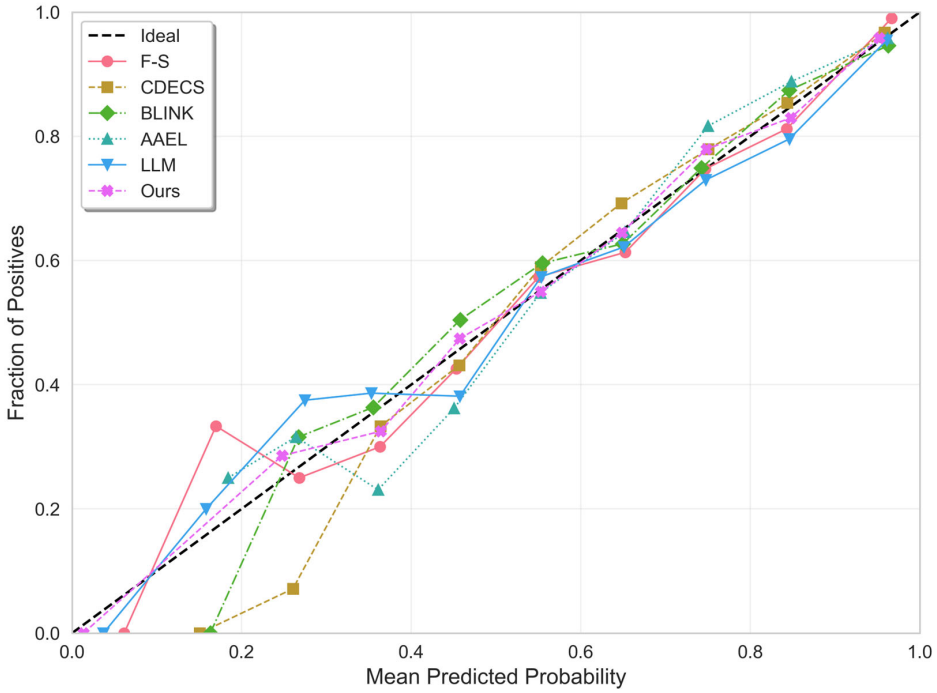

Figure 6. Probability calibration curves for all models, comparing mean predicted probability (x-axis) against empirical accuracy (y-axis). The proposed method (Ours) most closely follows the ideal calibration line (black dashed), demonstrating superior reliability in its confidence scores. The curves for baseline models reveal systematic miscalibration: for instance, F-S consistently overestimates probabilities, while others show inconsistent errors across confidence levels.

5.2. Stability Analysis Across Different Sources

To further evaluate the robustness of each model in practical settings, this subsection examines the impact of text source on model performance. We partition the test set into three categories by source: government portals (e.g., National Energy Administration, State Grid, Local Emergency Management Authorities), professional institutions (e.g., Polaris Power, Xuexi Qiang'an integrated media platform), and public media (e.g., NetEase News, Chudian News). We report Hit@1 and F1 for each source, as shown in Table 2. For fairness, the candidate-set generation pipeline, probability calibration, and decision thresholds remain the same across sources (fixed from validation), and the test split is identical for all methods.

Table 2. Performance on texts from different sources (%).

Source	Metric	F-S	CDECS	BLINK	AAEL	LLM	Ours
Govt. portals	Hit@1	42.21	40.33	38.17	43.56	41.82	42.97
	F1	59.87	60.25	57.89	61.34	61.02	63.11
Prof. inst.	Hit@1	32.74	39.15	36.92	42.8	41.35	42.08
	F1	56.92	59.87	56.34	61.05	61.44	63.27
Pub. media	Hit@1	31.58	35.47	34.26	39.21	38.74	39.87
	F1	55.33	56.82	54.17	58.96	59.21	60.53

In the cross-source comparison, the Hit@1 curves exhibit pronounced layering and fluctuations. Specifically, all models perform best on government texts, followed by professional-institution texts, and drop to the lowest on public-media texts. As shown in Figure 7a, the curves of AAEL and our method (Ours) remain in the top tier with relatively smooth trajectories, indicating stronger cross-domain stability; in contrast, F-S shows the largest fluctuations, with a marked decline on media data, suggesting weaker adaptability to non-standardized texts. This ordering is consistent with the F1 pattern in Figure 7b and reflects increasing structural variability from government to public-media sources.

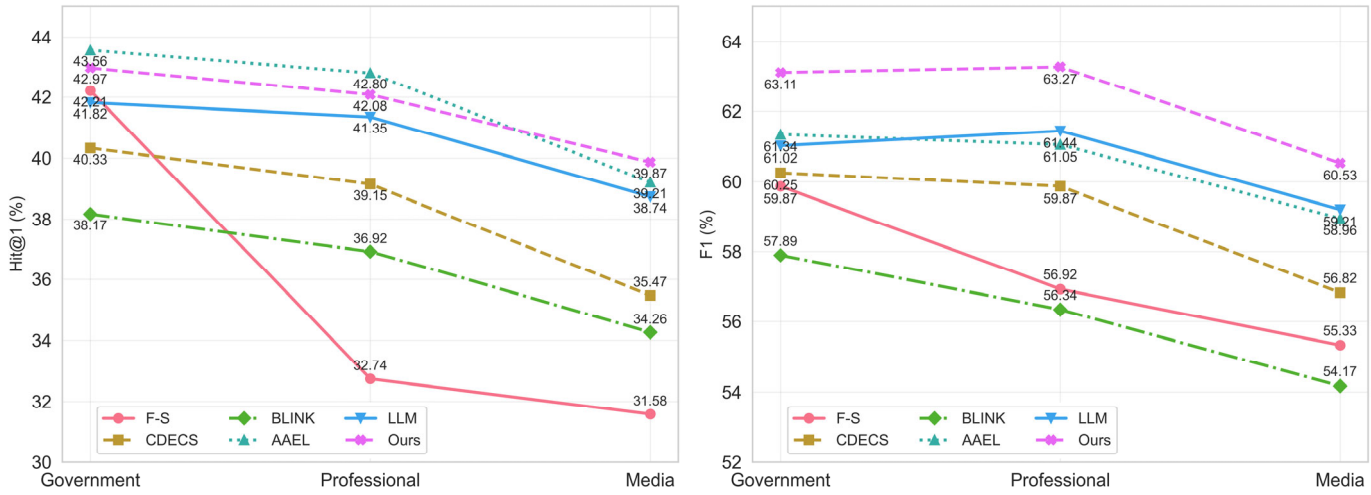

Figure 7. Model performance and robustness evaluation across three data sources: government portals, professional institutions, and public media. (**a**) Hit@1 performance comparison of six methods. (**b**) F1 score comparison of the same methods. The results reveal a consistent performance hierarchy (government > professional > media) across all models, reflecting increasing structural variability. The proposed method (Ours), along with AAEL, maintains top-tier performance with smooth trajectories, demonstrating superior cross-domain stability compared to baselines like F-S, which exhibit significant performance drops on less-structured media texts.

To better understand source-specific differences in information structure, we randomly sampled 100 accident reports from each source and computed the completeness of extracted accident information. We focused on coverage across nine key fields: time of occurrence, location, casualties, economic loss, involved project, involved organization, operational stage, accident type, and injury-causing agent. This sampling analysis is descriptive and is used solely to explain performance differences rather than to train or tune models. Figure 8 presents the coverage radar chart.



Figure 8. Sampling statistics of information coverage across sources. The uniform "fan" for government sources reflects systematic reporting standards; the locally enlarged sectors for professional institutions reveal technical emphasis; and the fragmented distribution for public media suggests selective reporting.

Figure 8 intuitively shows the differentiated coverage patterns. Government sources are most balanced, with broad arcs on basic fields (time, location, casualties) but no disclosure for involved organization, involved project, and injury-causing agent, consistent with standardized reporting requirements. Such structurally regular texts—with fixed information positions—benefit rule-based or traditional feature-matching models (e.g., F-S), which can achieve higher accuracy on basic fields. The lack of certain fields is largely attributable to privacy/compliance practices in public reporting, which reduces the availability of identifiers for linking.

By contrast, professional-institution texts exhibit prominent sectors on operational stage, accident type, and injury-causing agent, highlighting a focus on technical analysis. Terminology-dense, detail-rich documents favor models with semantic understanding (e.g., CDECS, LLM), which capture technical attributes and therefore perform well on cause- and agent-related fields. Traditional keyword models tend to degrade under such semantic complexity. Our method, which integrates spatiotemporal consistency with calibrated semantic fusion, maintains competitive Hit@1 and the best F1 in this source (Table 2), indicating robustness to terminology-dense narratives.

Public-media texts cover basic information like time and location reasonably well, but the sectors for involved organization and injury-causing agent narrow significantly, reflecting timeliness and selectivity in news reporting. Their looser structure, redundant

details, and narrative style pose challenges to all models. Even so, the high coverage of basic fields provides surface-matching opportunities, enabling models with stronger semantic generalization (e.g., Ours) to maintain relatively stable overall performance. Consistent with Table 2, our method attains the best F1 (60.53%) and the top Hit@1 (39.87%) on public-media texts, suggesting that calibrated probabilities help resist noise introduced by narrative redundancy.

In sum, accident texts from different sources exhibit systematic structural differences that markedly affect model performance. Government texts are standardized and field-complete, favoring traditional matching models but limiting deep reasoning due to avoidance of sensitive details. Professional-institution texts are terminology-dense with rich technical detail, better suited to semantically capable models for deeper analysis and linking. Public-media texts are loosely structured and information-redundant, challenging for all models, yet their high coverage of basic fields still benefits models with robust generalization. The findings indicate that models relying on surface feature matching generalize poorly across sources, whereas models incorporating semantic understanding and domain adaptation better accommodate structural differences and thus maintain more stable cross-source performance. From a deployment perspective, these observations motivate reporting both ranking/classification and calibration metrics by source so that operating thresholds for auto-labeling vs. human review can be tuned to each source's structural profile.

5.3. Stability Analysis Under Different Decision Thresholds

In practical deployments of event linking, the choice of decision threshold directly affects both predictive outcomes and utility. To comprehensively assess each model's stability across thresholds, we examine how the F1 score balance varies under different operating points, with the aim of verifying whether our calibration advantage yields more reliable outputs across settings. The threshold also defines the operational boundary between automatic labeling and human review, making stability across thresholds practically important.

To ensure comparability, we use a unified test set and fixed model parameters and evaluate three decision thresholds—0.3, 0.5, and 0.7—computing F1 for each model at each threshold while keeping all other experimental conditions unchanged. Temperature-scaling parameters for probability calibration remain fixed from the validation stage to avoid leakage and to ensure a fair comparison. The three thresholds correspond to lenient (0.3), default (0.5), and conservative (0.7) operating points. Table 3 and Figure 9 report the F1 scores at each threshold and the across-threshold fluctuation range (Δ).

Table 3. F1 scores at different decision thresholds (%).

Model	Threshold = 0.3	Threshold = 0.5	Threshold = 0.7	Range (Δ)
F-S	55.26	57.74	53.83	−3.91
CDECS	56.83	58.79	55.35	−3.44
BLINK	53.43	56.26	51.74	−4.52
AAEL	58.97	60.52	57.27	−3.25
LLM	59.13	60.71	56.97	−3.74
Ours	61.25	62.56	60.82	−1.74

Our method maintains consistently high F1 across thresholds and exhibits the smallest fluctuation (only 1.74%), indicating the best threshold robustness. In contrast, baselines fluctuate between 3.25% and 4.52% (AAEL 3.25, CDECS 3.44, LLM 3.74, F-S 3.91, BLINK 4.52), suggesting greater sensitivity to threshold selection. This aligns with our calibration advantage (lower Brier/ECE in Section 5.1), which reduces the mismatch between

predicted probabilities and empirical accuracy and thus dampens metric oscillation across operating points.

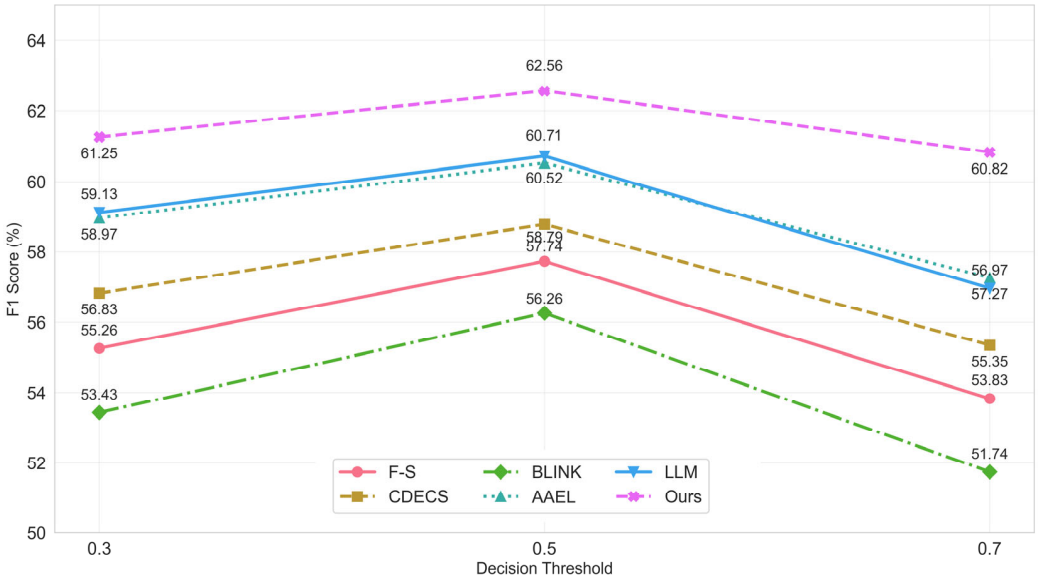

Figure 9. Threshold robustness analysis: F1 score stability of each model under different decision thresholds (0.3, 0.5, 0.7). The proposed method (Ours) not only achieves the highest F1 score at every threshold but also exhibits the smallest performance fluctuation ($\Delta = 1.74\%$), demonstrating superior insensitivity to operational point selection. This stability, a direct benefit of its well-calibrated probabilities, is critical for reliable system deployment.

5.4. Automatic Labeling Efficacy Analysis

We evaluate automatic labeling efficacy from an active-learning perspective. The analysis is based on the "maximum probability + probability gap" fusion strategy, which identifies high-uncertainty cases and precisely delineates the boundary between auto-labeling and human review. Metrics include the proportion of auto-labeled items, the proportion sent to human review, the accuracy of high-confidence auto-labels, and the proportion of true positives within the human review set. The execution results are shown in Table 4. Following the protocol in Section 3.5, both the decision threshold and the probability-gap parameter are fixed on the validation set and held constant on the test set to prevent leakage and enable fair comparison. For clarity, Auto denotes the share of texts automatically labeled; Acc (HC auto) is the accuracy among high-confidence auto-labels; Review is the share routed to human assessment; and TPR (review set) is the true-positive rate within the items sent to review. This design aligns with the PPRL triage principle by separating high-confidence, low-risk cases for automation from ambiguous cases that require human scrutiny.

Table 4. Comparison of automatic labeling efficiency across models (%).

Model	Auto	Acc (HC Auto)	Review	TPR (Review Set)
F-S	58.8	92.37	41.2	68.52
CDECS	67.5	94.11	32.5	72.36
BLINK	62.6	93.27	37.4	70.14
AAEL	70.2	95.45	29.8	75.67
LLM	71.4	95.83	28.6	76.22
Ours	**77.7**	**97.51**	**22.3**	**81.46**

The results indicate that our method delivers the best automatic labeling efficacy. The auto-label proportion reaches 77.7%, while the human review proportion drops to 22.3%. The accuracy of high-confidence auto-labels is 97.51%, indicating highly reliable automated outputs; within the human review set, the true-positive rate is 81.46%, confirming that the strategy effectively routes difficult, high-value cases for manual inspection. Compared with the F-S baseline (auto-label 58.8%, human review 41.2%), our approach reduces the human review share by 18.9 percentage points. Competing models (CDECS, BLINK, AAEL, LLM) achieve auto-label shares between 62.6% and 71.4%, all below our method. Furthermore, Acc (HC auto) of 97.51% exceeds all baselines by 1.68–5.14 pp (vs. 95.83% for the strongest competitor and 92.37% for F-S), while the review-set TPR of 81.46% indicates that the human queue is enriched for genuinely positive cases—an operationally desirable triage property. These advantages stem from more accurate probability calibration coupled with the "maximum-probability + probability-gap" criterion, which together provide reliable uncertainty estimates and a principled boundary between automated processing and human review, thereby improving throughput and quality while avoiding rework from erroneous labels.

5.5. Ablation Study

To evaluate the contribution of each proposed component, an ablation study was performed by systematically removing one module at a time from the complete model while keeping all other settings, datasets, and hyperparameters unchanged. The evaluation followed the same experimental protocol as described in Section 4, and all results were reported on the identical test set to ensure fair comparison. The components analyzed include:

(i) Spatiotemporal lattice consistency module (ϕ_S, ϕ_T), which provides spatial–temporal constraints using the RCC-8 and Allen-subset relations;
(ii) Smoothing and monotonic projection module, which regularizes weight estimation and preserves the natural order among relation strengths;
(iii) Semantic compensation term (ϕ_{sem}), which enhances robustness to linguistic variation;
(iiii) Temperature-based probability calibration, which improves the reliability of probabilistic outputs.

Each ablation variant was retrained from scratch to isolate its effect. Table 5 summarizes the results in terms of F1 (linking accuracy), Brier score (probability error), and ECE (expected calibration error).

Table 5. Ablation results on the main task.

Model Variant	F1 (%)	ΔF1 (%)	Brier	ΔBrier	ECE (%)	ΔECE (%)
Full Model	62.46	–	0.14	–	1.97	–
w/o ϕ_S, ϕ_T	57.21	−5.25	0.18	0.04	3.84	1.87
w/o ϕ_{sem}	60.28	−2.18	0.15	0.01	2.22	0.25
w/o Projection	59.04	−3.42	0.16	0.02	2.91	0.94
w/o Calibration	62.11	−0.35	0.17	0.03	4.08	2.11

The full model achieves the best overall performance (F1 = 62.46%, Brier = 0.14, ECE = 1.97%). Removing the spatiotemporal lattice consistency causes the most severe degradation, reducing F1 to 57.21% (−5.25%) and worsening calibration (Brier = 0.18, ECE = 3.84%). This confirms that the RCC-8/Allen-based lattice is fundamental for interpreting anonymized quasi-identifiers and ensuring coherent matching decisions.

Eliminating the semantic compensation term ϕ_{sem} results in a smaller but consistent decrease in recall, leading to F1 = 60.28% (−2.18%), Brier = 0.15, and ECE = 2.22%. The drop is most pronounced for narrative-heavy public-media sources, where anonymization

causes missing quasi-identifiers. This finding demonstrates that the semantic component complements lattice-based reasoning by restoring soft contextual clues.

Without smoothing and monotonic projection, the model exhibits less stable behavior, with F1 = 59.04% (−3.42%), Brier = 0.16, and ECE = 2.91%. This indicates that unsmoothed empirical weights introduce overconfidence and disrupt the monotonic relationship among spatial and temporal relations, impairing calibration stability.

Finally, removing temperature-based calibration barely affects F1 (62.11%, −0.35%) but substantially degrades probability reliability (Brier = 0.17, ECE = 4.08%). The absence of calibration results in overconfident scores and higher threshold sensitivity, which undermines auditability in privacy-preserving record linkage. The ablation results highlight three main insights:

(i) Spatiotemporal consistency is indispensable for reasoning over anonymized accident reports, enabling structured correlation of location and time even when explicit identifiers are removed.
(ii) Statistical regularization via smoothing and monotonic projection significantly enhances stability and prevents overfitting to sparse relation frequencies.
(iii) Semantic compensation improves recall under heterogeneous textual expressions, while probability calibration ensures the model's predictions retain auditable probabilistic meaning.

Together, these modules jointly ensure both quantitative accuracy and qualitative interpretability, validating the necessity of each design choice for reliable, privacy-preserving record linkage.

6. Results

The proposed method delivers consistently strong results across ranking, classification, and probability-based metrics. On the test set, it attains Hit@1 of 41.51% and Hit@5 of 77.33%, while achieving the best F1 of 62.46% with Precision 73.92% and Recall 54.07%. Probability-based discrimination is also competitive (ROC-AUC 87.34%). Most importantly for deployment, we follow a validation-fit/test-fix protocol for probability calibration and thresholds. The model exhibits the most reliable probability estimates among all systems evaluated, with the lowest Brier score (0.14) and the lowest ECE (1.97%). These properties are crucial in privacy-preserving record linkage (PPRL) settings, where calibrated probabilities support auditable thresholding and governance. Together, the findings indicate that small differences in Hit@k do not fully capture downstream reliability, whereas improvements in F1 and calibration translate into more trustworthy confidence scores for operational decisions.

Stratified analyses show stable performance across heterogeneous sources, including government portals, professional institutions, and public media, with the expected ordering—best on government texts and worst on public-media texts—holding across all methods. Our approach remains in the top tier for each source and shows reduced variability relative to feature-matching baselines. This robustness is attributable to the combination of structured spatiotemporal constraints and semantic modeling: the RCC-8-based spatial relations and the Allen-subset temporal relations enforce interpretable containment consistency (ϕ_S, ϕ_T), while the semantic compensation term (ϕ_{sem}) recovers soft evidence when quasi-identifiers are generalized or omitted.

Threshold-sensitivity experiments further corroborate the method's stability. Sweeping the decision threshold over 0.3, 0.5, and 0.7, the model sustains the highest F1 at each operating point and exhibits the smallest fluctuation range (Δ = 1.74%) among all systems. In practice, such reduced sensitivity is consistent with better probability calibration and the use of constrained monotonic fusion, simplifying deployment scenarios where operating

points may change with workload or policy. Consequently, operating-point selection can be aligned with review budgets while maintaining consistent utility across regimes.

Finally, the automatic labeling study demonstrates clear practical gains. Using a decision rule that combines maximum probability with a probability-gap criterion, the system automatically labels 77.7% of cases with 97.51% accuracy among high-confidence outputs, while routing 22.3% to human review; within the review subset, the true-positive rate reaches 81.46%. Relative to a strong feature-matching baseline, this reduces the human review share by 18.9%, increasing throughput without sacrificing quality. These outcomes reflect a calibrated triage: high-confidence items are automated, whereas ambiguous or cluster-adjacent cases—identified via the probability gap—are preferentially escalated to review, enriching the reviewer queue with genuinely informative positives.

Overall, the results show that integrating spatiotemporal lattice constraints with monotonic probability fusion and explicit calibration yields a model that is not only accurate but also operationally reliable. The ablation study further substantiates these contributions by quantifying the role of lattice consistency, smoothing/monotonic projection, semantic compensation, and temperature-based calibration. The approach provides calibrated confidence suitable for principled thresholding, maintains robustness across sources and operating points, and enables scalable, auditable automatic labeling in real-world accident-corpus construction.

7. Discussion

Our results confirm that explicit spatiotemporal containment modeling and end-to-end probability calibration jointly yield both competitive accuracy and operational reliability for cross-document accident linking. Concretely, the method matches strong baselines on ranking and discrimination while delivering the lowest Brier and lowest ECE, and it sustains the highest F1 across decision thresholds with the smallest performance fluctuation. These properties are especially valuable under PPRL constraints for strictly anonymized accident texts, where auditable probabilities and consistent behavior across operating points are prerequisites for deployment. In practice, this translates into an auto-label share of 77.7% with 97.51% accuracy on high-confidence outputs and a focused review queue (TPR 81.46%).

Viewed against prior lines of research, the approach complements both rule/template and graph-based methods—known for interpretability and global consistency but sensitive to similarity heuristics—as well as neural and PLM/LLM paradigms that excel at semantic alignment but can struggle when anchors are obscured by anonymization. By encoding administrative hierarchies and temporal granularity as a lattice (RCC-8/Allen subset) and employing a monotone combiner, the method injects domain knowledge that constrains plausible matches, thus mitigating lexical variability and granularity mismatch that frequently degrade open-domain systems when transferred to safety texts. This design reconciles two desiderata often viewed as competing: (i) interpretable, auditable decision factors and (ii) transferable semantic representations robust to style and source shifts. Concretely, interpretability is enforced by the monotone lattice combiner—raising spatial/temporal consistency can only increase the link score—and by a hard-conflict guard that deterministically penalizes AFTER/DISJOINT temporal–spatial contradictions before calibration; both signals are logged for audit. Semantic robustness then enters as a bounded compensation term: on the most challenging public-media texts, this balance yields the best F1 (60.53%) with top-tier Hit@1 (39.87%), showing that ϕ_{sem} can recover matches when quasi-identifiers are generalized, while the lattice prevents semantically plausible but topologically inconsistent links. Finally, explicit calibration converts constrained scores into trustworthy probabilities (Brier = 0.14; ECE = 1.97%), so borderline items with small

probability gaps are triaged to review rather than auto-labeled—consistent with our 22.3% review share enriched with true positives (81.46%).

The stratified analyses across government portals, professional institutions, and public media situate the results in a broader context of source-dependent structure. Government reports are standardized and field-complete, favoring all systems but especially those relying on surface features; professional-institution texts are terminology-dense and analytically rich, rewarding models with stronger semantic understanding; public-media reports are loosely structured and selective, challenging every method. Our system's cross-source stability is best explained by its dual reliance on (a) lattice-based containment consistency, which regularizes inference when key fields are present, and (b) calibrated probabilities, which reduce overconfidence when fields are missing or conflicting—thereby preserving throughput without inflating false positives.

Threshold-sensitivity experiments illuminate how calibration informs governance in human-in-the-loop pipelines. The two-tier rule—requiring both a minimum probability and a probability gap—operationalizes a conservative policy: obvious cases are automated; ambiguous, cluster-adjacent items are triaged for review. Unlike systems that implicitly conflate similarity with probability, our framework explicitly calibrates outputs (via temperature scaling) and logs decision metadata (strength/gap, relation categories, penalties), enabling principled choices of operating points tied to desired review budgets or service-level constraints. This aligns with deployment needs in critical domains, where auditability matters as much as mean accuracy.

At the same time, the study reveals limitations that motivate further work. First, upstream extraction errors (e.g., time normalization, location resolution, argument parsing) propagate into fusion scores; while metadata logging and spot checks mitigate this, extraction quality remains a bottleneck. Second, small-sample calibration with monotonic projection may underrepresent rare factor combinations; when the calibration set lacks coverage, probability estimates can be conservative or biased in the tails. Third, the fixed conflict penalty may over-suppress true positives in edge cases where administrative boundaries and reporting practices diverge (e.g., cross-jurisdiction incidents), suggesting a need for context-aware penalty schedules. In addition, feature sparsity in long-tail categories can limit the reliability of categorical compatibilities (ϕ_E); hierarchical smoothing or Bayesian shrinkage may further stabilize these estimates.

These observations point to several future directions: (1) Joint extraction–linking with uncertainty propagation (e.g., marginalizing over NER/normalization hypotheses) could reduce error cascades. (2) Adaptive calibration—such as hierarchical temperature scaling or Bayesian calibration—may better handle long-tail regimes and data drift. (3) Open-set detection and near-duplicate clustering could curb false links in bursty incident streams. (4) Multilingual and cross-regional extensions would test the lattice under diverse administrative ontologies and reporting norms. (5) Human-in-the-loop interfaces that expose calibrated probabilities, gap margins, and factor-wise contributions could improve reviewer efficiency and trust. (6) Finally, releasing auditable benchmarks with structured–unstructured pairs and calibration targets would facilitate apples-to-apples comparisons and accelerate progress on reliable event linking in safety-critical domains. Where feasible, these directions should preserve PPRL constraints by avoiding direct identifiers and documenting any relaxation via explicit governance policies.

Recent safety engineering research underscores the importance of analyzing spatiotemporal patterns to prevent accidents. Li et al. [57] mined 174 coal mine explosion investigation reports, categorized them by time and region and applied frequent itemset mining to identify coordination failures and risk propagation paths. Xie et al. [58] analyzed spatial–temporal distributions and used correlation and regression models to trace causal

factors in urban underground pipeline disasters. These studies rely on summarizing existing reports rather than linking anonymized events. In contrast, our lattice-constrained framework goes further by encoding administrative hierarchies and temporal granularity, instantiating relations via RCC-8 and Allen's interval algebra, and calibrating probabilities. This cross-document event linking capability not only produces auditable matches but also enables construction of high-quality accident corpora, providing a technical foundation for data-driven accident prevention research. In sum, by fusing structured constraints with calibrated probabilistic inference, the proposed system advances beyond accuracy alone toward deployable reliability—maintaining performance across sources and thresholds, reducing manual burden, and providing transparent signals for governance and audit. This offers a concrete and extensible path for building large-scale, auditable accident corpora under privacy constraints, integrating interpretable lattice-based reasoning with well-calibrated, semantics-aware fusion.

Author Contributions: Conceptualization and methodology, W.Z. and W.T.; data processing and experimental analysis, W.Z. and Y.Z.; validation, B.Z. and W.T.; investigation, B.Z.; writing—original draft preparation, W.Z.; writing—review and editing, W.Z. and Y.Z.; visualization, W.Z. and Y.Z.; supervision and project administration, W.T. and D.Y.; funding acquisition, W.T. and D.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by [Shenzhen Science and Technology Program], grant number [KCXFZ20230731093902005].

Data Availability Statement: The dataset used in this study can be made available by the corresponding author upon reasonable request.

Acknowledgments: We would like to thank the anonymous reviewers for their valuable feedback and helpful suggestions, which have greatly improved the quality of this manuscript. Special thanks to the City Safety Risk Monitoring and Early Warning Emergency Management Key Laboratory at the Shenzhen City Public Safety Technology Research Institute for their technical support and for providing the necessary resources to conduct this research.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Gkoulalas-Divanis, A.; Vatsalan, D.; Karapiperis, D.; Kantarcioglu, M. Modern Privacy-Preserving Record Linkage Techniques: An Overview. *IEEE Trans. Inf. Forensics Secur.* **2021**, *16*, 4966–4987. [CrossRef]
2. Nothman, J.; Honnibal, M.; Hachey, B.; Curran, J.R. Event Linking: Grounding Event Reference in a News Archive. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Jeju, Republic of Korea, 8–14 July 2012; pp. 228–232.
3. Humphreys, K.; Gaizauskas, R.; Azzam, S. Event Coreference for Information Extraction. In Proceedings of the Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Text, Madrid, Spain, 11 July 1997; Association for Computational Linguistics: Stroudsburg, PA, USA, 1997; pp. 75–81.
4. Ahn, D. The Stages of Event Extraction. In *Workshop on Annotating and Reasoning About Time and Events*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2006; pp. 1–8.
5. Raghunathan, K.; Lee, H.; Rangarajan, S.; Chambers, N.; Surdeanu, M.; Jurafsky, D.; Manning, C.D. A Multi-Pass Sieve for Coreference Resolution. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Cambridge, MA, USA, 9–11 October 2010.
6. Lu, J.; Ng, V. Event Coreference Resolution with Multi-Pass Sieves. In Proceedings of the Tenth International Conference on Language Resources and Evaluation, Portorož, Slovenia, 23–28 May 2016; pp. 3996–4003.
7. Cybulska, A.; Vossen, P. Using a Sledgehammer to Crack a Nut? Lexical Diversity and Event Coreference Resolution. In Proceedings of the Ninth International Conference on Language Resources and Evaluation, Reykjavik, Iceland, 26–31 May 2014; pp. 4545–4552.

8. Cybulska, A.; Vossen, P. "Bag of Events" Approach to Event Coreference Resolution: Supervised Classification of Event Templates. *Int. J. Comput. Linguist. Appl.* **2015**, *6*, 11–27.
9. Cybulska, A.; Vossen, P. Translating Granularity of Event Slots into Features for Event Coreference Resolution. In Proceedings of the 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation, Denver, CO, USA, 4 June 2015; pp. 1–10.
10. Hovy, E.; Mitamura, T.; Verdejo, F.; Araki, J.; Philpot, A. Events Are Not Simple: Identity, Non-Identity, and Quasi-Identity. In Proceedings of the 1st Workshop on Events: Definition, Detection, Coreference, and Representation, Atlanta, GA, USA, 14 June 2013; pp. 21–28.
11. Araki, J.; Hovy, E.; Mitamura, T. Evaluation for Partial Event Coreference. In Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation, Baltimore, MD, USA, 22–27 June 2014; pp. 68–76.
12. Chen, Z.; Ji, H. Graph-Based Event Coreference Resolution. In Proceedings of the 2009 Workshop on Graph-Based Methods for Natural Language Processing, Singapore, 7 August 2009; pp. 54–57.
13. Chen, Z.; Ji, H.; Haralick, R. A Pairwise Event Coreference Model, Feature Impact and Evaluation for Event Coreference Resolution. In Proceedings of the Workshop on Events in Emerging Text Types (eETTs) 2009, Borovets, Bulgaria, 17 September 2009; pp. 17–22.
14. Lee, H.; Recasens, M.; Chang, A.; Surdeanu, M.; Jurafsky, D. Joint Entity and Event Coreference Resolution across Documents. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Jeju Island, Republic of Korea, 12–14 July 2012; pp. 489–500.
15. Choubey, P.K.; Huang, R. Event Coreference Resolution by Iteratively Unfolding Inter-Dependencies among Events. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 2124–2133.
16. Yang, B.; Cardie, C.; Frazier, P. A Hierarchical Distance-Dependent Bayesian Model for Event Coreference Resolution. *Trans. Assoc. Comput. Linguist.* **2015**, *3*, 517–528. [CrossRef]
17. Kenyon-Dean, K.; Cheung, J.C.K.; Precup, D. Resolving Event Coreference with Supervised Representation Learning and Clustering-Oriented Regularization. In Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, New Orleans, LA, USA, 5–6 June 2018; pp. 1–10.
18. Peng, H.; Song, Y.; Roth, D. Event Detection and Co-Reference with Minimal Supervision. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 392–402.
19. Barhom, S.; Shwartz, V.; Eirew, A.; Bugert, M.; Reimers, N.; Dagan, I. Revisiting Joint Modeling of Cross-Document Entity and Event Coreference Resolution. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 4179–4189.
20. Cremisini, A.; Finlayson, M.A. New Insights into Cross-Document Event Coreference: Systematic Comparison and a Simplified Approach. In Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events, Online, 9 July 2020; pp. 7–16.
21. Allaway, E.; Wang, S.; Ballesteros, M. Sequential Cross-Document Coreference Resolution. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online, 7–11 November 2021; pp. 4659–4671.
22. Han, R.; Peng, T.; Yang, C.; Wang, B.; Liu, L.; Wan, X. Is Information Extraction Solved by ChatGPT? An Analysis of Performance, Evaluation Criteria, Robustness and Errors. *arXiv* **2023**, arXiv:2305.14450. [CrossRef]
23. Nath, A.; Avari, S..; Chelle, A.; Krishnaswamy, N. Okay, Let's Do This! Modeling Event Coreference with Generated Rationales and Knowledge Distillation. *arXiv* **2024**, arXiv:2404.03196.
24. Wang, X.; Zhou, W.; Zu, C.; Xia, H.; Chen, T.; Zhang, Y.; Zheng, R.; Ye, J.; Zhang, Q.; Gui, T.; et al. InstructUIE: Multi-Task Instruction Tuning for Unified Information Extraction. *arXiv* **2023**, arXiv:2304.08085.
25. Wolter, F.; Zakharyaschev, M. Spatio-Temporal Representation and Reasoning Based on RCC-8. In Proceedings of the Seventh International Conference on Principles of Knowledge Representation and Reasoning, Breckenridge, CO, USA, 11–15 April 2000; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2000; pp. 3–14.
26. Allen, J.F. Maintaining Knowledge about Temporal Intervals. *Commun. ACM* **1983**, *26*, 832–843. [CrossRef]
27. Cybulska, A.; Vossen, P. *Guidelines for ECB+ Annotation of Events and their Coreference*; Technical Report; VU University Amsterdam: Amsterdam, The Netherlands, 2014.
28. Ellis, J.; Getman, J.; Fore, D.; Kuster, N.; Song, Z.; Bies, A.; Strassel, S.M. Overview of Linguistic Resources for the TAC KBP 2015 Evaluations: Methodologies and Results. In Proceedings of the TAC 2015, Gaithersburg, MD, USA, 16–17 November 2015; National Institute of Standards and Technology (NIST): Gaithersburg, MD, USA, 2015.
29. Chen, B.; Su, J.; Pan, S.J.; Tan, C.L. A Unified Event Coreference Resolution by Integrating Multiple Resolvers. In Proceedings of the 5th International Joint Conference on Natural Language Processing, Chiang Mai, Thailand, 8–13 November 2011; pp. 102–110.

30. Lu, J.; Venugopal, D.; Gogate, V.; Ng, V. Joint Inference for Event Coreference Resolution. In Proceedings of the COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan, 11–16 December 2016; pp. 3264–3275.
31. Lu, J.; Ng, V. Joint Learning for Event Coreference Resolution. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 90–101.
32. Bejan, C.A.; Harabagiu, S. Unsupervised Event Coreference Resolution with Rich Linguistic Features. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010; Association for Computational Linguistics: Stroudsburg, PA, USA, 2010; pp. 1412–1422.
33. Bejan, C.A.; Harabagiu, S. Unsupervised Event Coreference Resolution. *Comput. Linguist.* **2014**, *40*, 311–347. [CrossRef]
34. Araki, J.; Mitamura, T. Joint event trigger identification and event coreference resolution with structured perceptron. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 2074–2080.
35. Upadhyay, S.; Gupta, N.; Christodoulopoulos, C.; Roth, D. Revisiting the Evaluation for Cross-Document Event Coreference. In Proceedings of the COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan, 11–17 December 2016; pp. 1949–1960.
36. Joshi, M.; Levy, O.; Weld, D.S.; Zettlemoyer, L. BERT for Coreference Resolution: Baselines and Analysis. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, 3–7 November 2019; pp. 5803–5808.
37. Joshi, M.; Chen, D.; Liu, Y.; Weld, D.S.; Zettlemoyer, L.; Levy, O. Spanbert: Improving Pre-Training by Representing and Predicting Spans. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 64–77. [CrossRef]
38. Cattan, A.; Eirew, A.; Stanovsky, G.; Joshi, M.; Dagan, I. Streamlining Cross-Document Coreference Resolution: Evaluation and Modeling. *arXiv* **2020**, arXiv:2009.11032. [CrossRef]
39. Yu, X.; Yin, W.; Roth, D. Paired representation learning for event and entity coreference[OL]. *arXiv* **2010**, arXiv:2010.12808.
40. Zeng, Y.; Jin, X.; Guan, S.; Guo, J.; Cheng, X. Event Coreference Resolution with Their Paraphrases and Argument-Aware Embeddings. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 8–13 December 2020; pp. 1737–1747.
41. Caciularu, A.; Cohan, A.; Beltagy, I.; Peters, M.E.; Cattan, A.; Dagan, I. CDLM: Cross-Document Language Modeling. *arXiv* **2021**, arXiv:2101.00406. [CrossRef]
42. Beltagy, I.; Peters, M.E.; Cohan, A. Longformer: The Long-Document Transformer. *arXiv* **2020**, arXiv:2004.05150. [CrossRef]
43. Ma, Y.; Cao, Y.; Hong, Y.; Sun, A. Large Language Model Is Not a Good Few-Shot Information Extractor, but a Good Reranker for Hard Samples! In *Findings of the Association for Computational Linguistics: EMNLP 2023*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 10572–10601.
44. Li, J.; Jia, Z.; Zheng, Z. Semi-Automatic Data Enhancement for Document-Level Relation Extraction with Distant Supervision from Large Language Models. *arXiv* **2023**, arXiv:2311.07314.
45. Min, Q.; Guo, Q.; Hu, X.; Huang, S.; Zhang, Z.; Zhang, Y. Synergetic Event Understanding: A Collaborative Approach to Cross-Document Event Coreference Resolution with Large Language Models. *arXiv* **2024**, arXiv:2406.02148. [CrossRef]
46. Ding, B.; Min, Q.; Ma, S.; Li, Y.; Yang, L.; Zhang, Y. A Rationale-Centric Counterfactual Data Augmentation Method for Cross-Document Event Coreference Resolution. *arXiv* **2024**, arXiv:2404.01921.
47. Bugert, M.; Reimers, N.; Gurevych, I. Generalizing Cross-Document Event Coreference Resolution across Multiple Corpora. *Comput. Linguist.* **2021**, *47*, 575–614. [CrossRef]
48. Field, D.A. Laplacian Smoothing and Delaunay Triangulations. *Commun. Appl. Numer. Methods* **1988**, *4*, 709–712. [CrossRef]
49. Yuan, Q.; Cong, G.; Thalmann, N.M. Enhancing Naive Bayes with Various Smoothing Methods for Short Text Classification. In Proceedings of the 21st International Conference on World Wide Web, Lyon, France, 16–20 April 2012; pp. 645–646.
50. Neelon, B.; Dunson, D.B. Bayesian Isotonic Regression and Trend Analysis. *Biometrics* **2004**, *60*, 398–406. [CrossRef]
51. De Leeuw, J.; Hornik, K.; Mair, P. Isotone Optimization in R: Pool-Adjacent-Violators Algorithm (PAVA) and Active Set Methods. *J. Stat. Softw.* **2010**, *32*, 1–24. [CrossRef]
52. Fellegi, I.P.; Sunter, A.B. A Theory for Record Linkage. *J. Am. Stat. Assoc.* **1969**, *64*, 1183–1210. [CrossRef]
53. Eirew, A.; Caciularu, A.; Dagan, I. Cross-Document Event Coreference Search: Task, Dataset and Modeling. *arXiv* **2022**, arXiv:2210.12654. [CrossRef]
54. Wu, L.; Petroni, F.; Josifoski, M.; Riedel, S.; Zettlemoyer, L. Scalable Zero-Shot Entity Linking with Dense Entity Retrieval. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 6397–6407.

55. Hsu, I.-H.; Xue, Z.; Pochhi, N.; Bansal, S.; Natarajan, P.; Srinivasa, J.; Peng, N. Argument-Aware Approach To Event Linking. In *Findings of the Association for Computational Linguistics: ACL 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 12769–12781.
56. Kim, T.; Oh, J.; Kim, N.Y.; Cho, S.; Yun, S.-Y. Comparing Kullback–Leibler Divergence and Mean Squared Error Loss in Knowledge Distillation. *arXiv* **2021**, arXiv:2105.08919.
57. Li, W.; Du, G.; Chen, L.; Zhang, R.; Chen, A. A Spatial-Temporal Exploration of Coordination Failures Preceding Coal Mine Explosion Accidents in China. *Sustainability* **2024**, *17*, 85. [CrossRef]
58. Xie, Z.; Jiang, F.; Xu, J.; Zhai, Z.; He, J.; Zheng, D.; Lian, J.; Hou, Z.; Zhao, L.; Wang, Y.; et al. A Narrative of Urban Underground Pipeline System Disasters in China in 2021: Spatial and Temporal Distribution, Causal Analysis, and Response Strategies. *Sustainability* **2023**, *15*, 10067. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.