

Article

HRE-LSC: A Hyper-Relational Data Enhancement Framework for Long Tail Distribution and Structural Consistency

Xinping Du ¹, Yan Chen ¹, Siqi Xu ¹ , Shuli Dong ¹ and Ping Feng ^{1,2,3,*} 

¹ College of Computer Science and Technology, Changchun University, Changchun 130022, China; xiaodu561@gmail.com (X.D.); potato66669999@gmail.com (Y.C.); xusiqi29@gmail.com (S.X.); dongshuli103@gmail.com (S.D.)

² Jilin Provincial Key Laboratory of Manchurian Railway Data Engineering and Intelligent Decision-Making, Changchun 130022, China

³ Education Key Laboratory of Intelligent Rehabilitation and Barrier-Free Access for the Disabled, Changchun University, Changchun 130022, China

* Correspondence: pingfeng20188@gmail.com

Abstract

Hyper-relational extraction aims to identify complex factual structures from unstructured text, where each instance consists of a core triplet and multiple qualified attributes. However, existing hyper-relational datasets often suffer from long-tail relation distributions and pseudo-negative samples, which limit the performance of hyper-relational extraction models. Although large language models (LLMs) provide new opportunities for data augmentation, they may introduce semantically inconsistent relations and structurally unreliable samples in complex scenarios, reducing the quality of generated data. To address these challenges, this paper proposes a structure-consistency-driven semi-automatic data augmentation framework, termed HRE-LSC. The framework consists of two key components: (1) a distribution-aware generation strategy that selectively generates samples for low-frequency relations according to relation frequency distributions, thereby alleviating data imbalance; and (2) a hierarchical logical consistency verification mechanism based on natural language inference (NLI), which progressively verifies the consistency of core triplets and qualified attributes to filter unreliable generated samples and reduce pseudo-negative samples. Experiments on the HyperRED dataset demonstrate that the proposed method improves the F1 score of Text2NKG from 83.81% to 84.7%, with more significant improvements observed on long-tail relation subsets. The results indicate that the proposed framework effectively enhances the quality and structural consistency of generated hyper-relational data while mitigating the effects of long-tail distributions and pseudo-negative samples without requiring additional manual annotations.

Keywords: hyper-relational extraction; long-tail distribution; large language models; hierarchical verification; natural language inference



Received: 22 June 2026

Revised: 12 July 2026

Accepted: 18 July 2026

Published: 21 July 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Knowledge graphs are progressively evolving from simple structures denoting mere binary relations to hyper-relational representations capable of characterizing rich contextual information [1]. In a hyper-relational paradigm, a triplet comprises not only a subject, a relation, and an object, but can also be augmented with qualifier attributes—such as time, location, and conditions—to delineate complex real-world facts with greater granularity [2]. For instance, the factual statement “Barack Obama served as the President of the United

States” necessitates the integration of supplementary attributes, such as “Time: 2008” or “Political Party: Democratic Party,” for a comprehensive articulation. Consequently, accurate hyper-relational extraction is of critical significance for downstream tasks, including complex knowledge reasoning, multi-hop question answering, and dynamic knowledge modeling [3,4].

However, existing hyper-relational extraction datasets encounter two primary challenges in practical applications [5]. First, the HyperRED dataset exhibits a pronounced long-tail distribution: a minority of head relations dominate the sample space, whereas numerous low-frequency relations and their corresponding attribute combinations remain critically scarce. This scarcity impedes the model’s ability to learn stable structural patterns, thereby constraining its generalization capabilities in complex scenarios [6]. As illustrated in Figure 1, the relation frequency distribution of the HyperRED dataset follows a typical long-tail pattern, where a small number of relations account for a large proportion of samples, while most low-frequency relations contain relatively limited instances. Second, owing to the prohibitive costs associated with manual annotation, datasets prevalently harbor unannotated implicit attributes, yielding pseudo-negative samples. The existing literature has proposed various paradigms to mitigate the ensuing training biases [7]. For instance, Xie investigated the quality of negative samples within the context of distantly supervised relation extraction, re-evaluating the pseudo-negative conundrum and demonstrating that certain negative samples inherently contain unannotated positive intelligence, which consequently compromises model training [8]; conversely, Zeng proposed a methodology for document-level relation extraction that integrates pseudo-negative mining with positive-class knowledge distillation to rectify erroneous annotations and amplify positive signals, thereby alleviating noise-induced detriments [9]. Nevertheless, these two categories of challenges are conventionally addressed in isolation, lacking a unified optimization framework, which ultimately impedes further advancements in the performance of hyper-relational extraction (HRE). The generative proficiency of large language models (LLMs) offers a prospective avenue for resolving both the long-tail and pseudo-negative dilemmas [10]. However, within the highly structured domain of hyper-relational extraction tasks, generative outputs are acutely susceptible to relation drift and semantic inconsistencies; this vulnerability is particularly exacerbated in multi-attribute compositional scenarios, wherein the direct assimilation of generated data may inadvertently inject noise, thereby stalling enhancements in model efficacy [11].

To address these impediments, this study proposes HRE-LSC, a hierarchical structure-consistency-driven hyper-relational data augmentation framework. The proposed framework executes targeted augmentation for long-tail relations via distribution-aware generation and formalizes the structural alignment of the generative outputs as a natural language inference (NLI) task. Specifically, it constructs a comprehensive permutation hypothesis space and incorporates an “entailment-contradiction” differential scoring mechanism to rigorously filter principal triplets alongside their qualifier attributes. Ultimately, a globally optimal Top-1 alignment strategy is deployed to guarantee high precision and preclude semantic conflicts. Comprehensive experiments conducted on the HyperRED dataset demonstrate that HRE-LSC significantly elevates the overall F1 score without necessitating supplementary manual annotations, and it attains substantial performance gains on the long-tail relational subsets, thereby corroborating the efficacy of structural consistency constraints within data augmentation methodologies.

The primary contributions of this paper are summarized as follows:

- A distribution-aware data augmentation strategy is proposed, which models the frequency distribution of training data to guide large language models in generating samples for long-tail relations in a targeted manner.

- A hierarchical structure consistency verification framework, HRE-LSC, is designed, which unifies hyper-relational structural alignment as a natural language inference task and achieves high-precision filtering via full-permutation hypotheses and differential scoring.
- An augmented dataset, HyperRED+, is constructed, which generates structurally consistent and high-quality hyper-relational data without manual annotation, and demonstrates significant improvements in both overall performance and long-tail subsets.

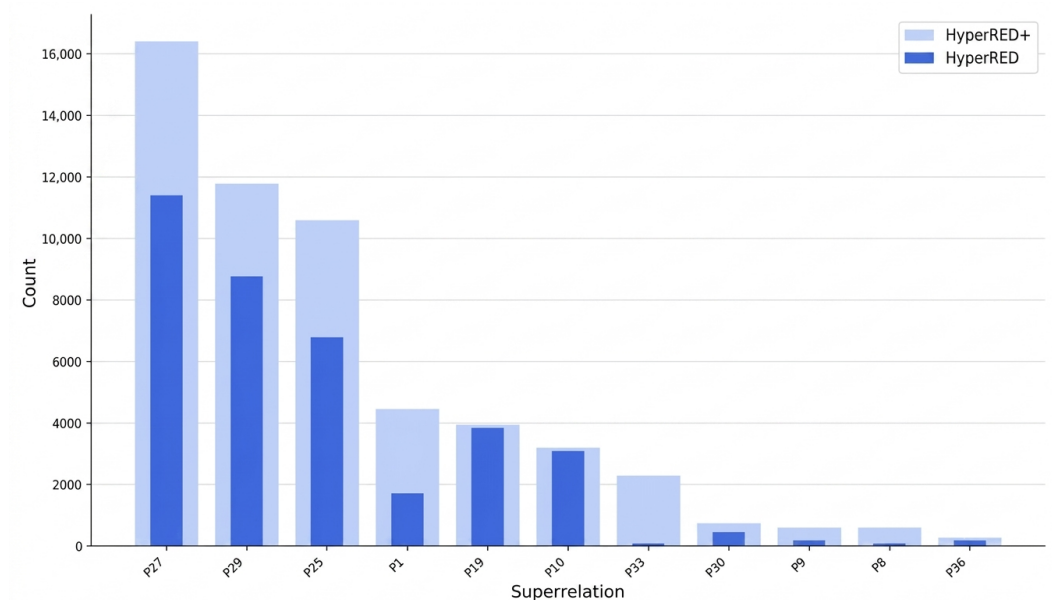


Figure 1. Statistics of the number of relation types across different datasets.

2. Related Work

Hyper-relational extraction endeavors to extract complex factual structures encompassing qualifier attributes from unstructured text; compared to traditional triplet-based extraction, the combinatorial space of relations and attributes is exponentially expanded [12]. Existing methodologies can be broadly bifurcated into two paradigms. The first category comprises pipeline and joint learning approaches. For example, Chia (2022) modeled the task as a cube-filling structure prediction problem, thereby achieving a unified representation of triplets and qualifier attributes [13], whereas Saini (2025) leveraged a multi-task learning framework to jointly model triplets and attributes to enhance structural consistency [14]. The second category encompasses generative methods; for instance, Luo (2024) reformulated hyper-relational extraction as a sequence generation task, facilitating the end-to-end generation of structured outputs and correspondingly diminishing reliance on handcrafted features [15]. Although the aforementioned approaches have achieved notable progress in structural modeling, their optimization trajectories predominantly focus on model architecture design, leaving critical data-level challenges inadequately addressed [16].

In complex structural scenarios, the long-tail distribution and pseudo-negative sample issues within training data are frequently tightly coupled, progressively emerging as the principal bottlenecks constraining model performance enhancements [17]. Specifically, a minority of high-frequency relations dominate real-world corpora, while low-frequency relations suffer from data scarcity, impeding the acquisition of stable representations. The existing literature has attempted to mitigate this long-tail dilemma from varying perspectives [18]. For example, Qiu (2023) modeled label dependencies and incorporated a multi-classifier ensemble mechanism to facilitate knowledge transfer between head and

tail categories, thereby elevating the recognition capability for low-frequency relations [19]. Tang (2023) introduced adaptive adjustment strategies and causal inference techniques to attenuate spurious correlations induced by data distribution biases, subsequently augmenting the generalization performance of long-tail relations [20]. Meng (2023) proposed a relation-aware prototype learning approach to address data scarcity in document-level relation extraction by constructing instance-level prototypes and relation-weighted contrastive learning [21]. Furthermore, Menotti (2026) proposed a long-tail-aware optimization framework tailored for document-level relation extraction to improve the recognition efficacy of low-frequency relations within complex contexts [22]. However, these methodologies are predominantly formulated for conventional relation extraction tasks and consequently exhibit inherent limitations within hyper-relational scenarios. In contrast, under hyper-relational conditions, the permutation of relations and attributes exacerbates data sparsity, rendering low-frequency relations increasingly susceptible to noise perturbations. Furthermore, extant approaches typically isolate the treatment of long-tail issues from pseudo-negative anomalies, failing to provide a unified modeling framework.

Furthermore, from the perspective of data acquisition mechanisms, while distant supervision and weak annotation strategies effectively curtail manual annotation expenditures, they inevitably introduce unannotated factual constructs, thereby generating a substantial corpus of pseudo-negative samples [23]. Tan (2022) re-examined the composition of negative samples in distantly supervised relation extraction, elucidating that certain negative samples inherently contain latent positive information, thereby exposing the disruptive impact of pseudo-negative samples on model training [24]. Hao (2021) proposed an adversarial training-based methodology to explicitly model pseudo-negative samples, aiming to alleviate the concomitant biases [25]. Additionally, Chen (2021) integrated pseudo-negative mining with positive-class knowledge distillation in document-level relation extraction to rectify noisy labels and fortify positive class representations [26]. Chen (2023) further extended this line by designing consistent prototype learning to suppress noisy label interference in continual extraction settings [27]. Nonetheless, these strategies predominantly optimize the pseudo-negative anomaly in isolation, devoid of a unified modeling framework. Extant research primarily relies on positive-unlabeled learning or adversarial training to mitigate noise interference, yet the explicit modeling of structural consistency remains markedly deficient.

Building upon this foundation, commensurate with the rapid advancement of large language models (LLMs), generative data augmentation has progressively emerged as an active research frontier [28]. Diaz-Garcia (2025) systematically reviewed language model-based relation extraction techniques from a survey perspective, emphasizing that generative paradigms and prompt learning are crystallizing as pivotal developmental trajectories [29]. Li (2023) leveraged LLMs in conjunction with distant supervision signals for semi-automatic data augmentation to exponentially expand the training data scale for document-level relation extraction [30]. Zheng (2024) explored the use of LLMs as fine-grained relation extraction data augmentors, demonstrating that LLM-based generation can improve sample diversity while maintaining semantic relevance through task-oriented augmentation strategies [31]. Swarup (2025) conducted a data-centric feasibility study on LLM-based relation extraction and systematically analyzed the robustness of LLMs under different data characteristics, revealing that LLMs still face challenges when handling complex relation extraction scenarios [10]. Popovic (2025) utilized LLMs to generate diversified synthetic demonstrations for information extraction tasks, further highlighting the potential of generative models in constructing training resources [32]. Meanwhile, Josifoski (2023) constructed a prompt learning framework based on synthetic exemplars, realizing zero-shot or weakly supervised information extraction through in-context learning [33]. Neverthe-

less, in complex structural scenarios, these methods remain highly susceptible to semantic drift and structural inconsistency. Particularly within hyper-relational extraction tasks, LLMs frequently manifest factual hallucinations and relation drift. The absence of rigorous constraints on complex structures fundamentally renders the generated data inadequate for satisfying high-precision task requirements.

In summary, existing methodologies still exhibit substantial deficiencies in three primary dimensions: (1) the absence of a unified framework to conjointly model the coupling dynamics between low-frequency relations and pseudo-negative samples; (2) the inherent difficulty of generative approaches in guaranteeing structural consistency for complex hyper-relational facts; and (3) the lack of an efficacious synergistic mechanism orchestrating data augmentation and structural constraints.

To resolve these deficiencies, this study proceeds from the premise of structural consistency modeling, proposing a methodology that intricately integrates open-ended data generation with hierarchical logical verification. This approach concurrently ameliorates the long-tail dilemma and imposes explicit structural constraints, thereby realizing high-precision, strongly consistent hyper-relational extraction.

3. Method

3.1. Problem Formulation

Given a sentence-level hyper-relational dataset $D = \{(S_i, T_i)\}_{i=1}^N$, where S_i denotes the input sentence and T_i represents the corresponding set of hyper-relational facts, each fact consists of a core triplet $t = (e_1, r, e_2)$ and an attribute set $A = \{(a_k, v_k)\}$, where a_k denotes the attribute type and v_k denotes the attribute value.

The objective of this work is to construct an augmented dataset D^+ without introducing additional manual annotations, such that

- the long-tail distribution of relations and attributes is alleviated;
- the training noise caused by pseudo-negative samples is reduced;
- the structural consistency of generated data is guaranteed.

3.2. Method Overview

To achieve the above objectives, we propose the HRE-LSC framework, as illustrated in Figure 2. The core idea is to decouple open-ended data generation from structural consistency verification, and to perform high-confidence filtering through hierarchical logical constraints.

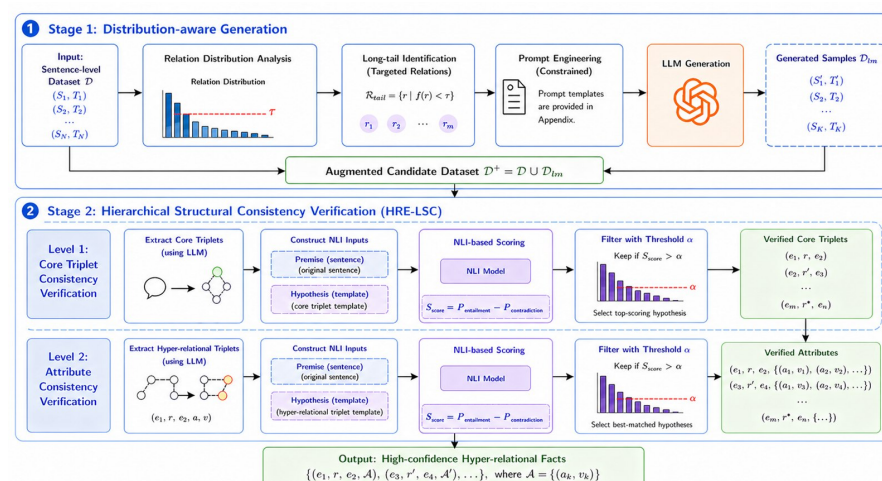


Figure 2. Overview of the proposed HRE-LSC framework.

The framework consists of two main stages:

- Distribution-aware generation, which performs targeted augmentation for long-tail relations;
- Hierarchical structural consistency verification, which filters high-quality samples from generated data.

The second stage is further divided into

- Level 1: Core triplet consistency verification;
- Level 2: Attribute consistency verification.

3.3. Distribution-Aware Data Generation Strategy

To alleviate the long-tail issue in the dataset, we first compute the frequency distributions of the relation set R and the attribute set A in the training data. Let the occurrence frequency of relation r be denoted as $f(r)$, and let τ denote the predefined frequency threshold for distinguishing long-tail relations. represent the predefined frequency threshold for identifying long-tail relations. A relation is considered long-tail when its occurrence frequency is lower than τ . Specifically, we define the long-tail relation set as:

$$R_{\text{tail}} = \{r \mid f(r) < \tau\} \quad (1)$$

In our experiments, τ is empirically set to 500 based on the frequency distribution of relations in the training set. This threshold is selected according to the following considerations: (1) relations with frequencies lower than τ represent insufficiently observed categories and are more likely to suffer from inadequate supervision during training; (2) selecting a larger threshold may introduce relatively frequent relations into the augmentation process, increasing unnecessary synthetic samples and potentially affecting the original data distribution; (3) selecting a smaller threshold may fail to cover severely under-represented relations. Therefore, τ provides a balance between improving the coverage of rare relations and maintaining the original distribution characteristics.

For each relation $r \in R_{\text{tail}}$, we construct a relation-specific generation strategy. Specifically, each long-tail relation is associated with a predefined hyper-relation attribute set A_r , which constrains the large language model to generate only valid hyper-relational facts corresponding to the target relation. The generation prompts explicitly provide the target relation r and its corresponding attribute set A_r , and require the generated samples to satisfy entity consistency, relation validity, and attribute compatibility constraints. The complete prompt templates used for LLM-based data generation are provided in Appendix B.

The generated samples are denoted as D_{llm} , and are combined with the original dataset to construct the augmented candidate set:

$$D^+ = D \cup D_{\text{llm}} \quad (2)$$

where D denotes the original dataset.

3.4. Hierarchical Structural Consistency Verification (HRE-LSC)

To address structural deviations introduced by large language models during open-ended generation, we propose a hierarchical structural consistency verification mechanism based on natural language inference (NLI). The key idea is to unify relation alignment and attribute matching in hyper-relational extraction as a structural consistency evaluation problem, and to progressively filter high-confidence facts through hierarchical constraints. The verification process consists of two levels: core triplet consistency verification and attribute consistency verification.

3.5. Level 1: Core Triplet Consistency Verification

Given a sentence S , we first use a large language model to generate candidate entity pairs and their potential relations, forming an initial candidate triplet set. To ensure sufficient coverage, no strict limit is imposed on the number of generated candidates, allowing exploration of diverse entity combinations and relation expressions. However, due to the lack of strict structural constraints in LLM outputs, the generated candidates may include relation type shifts or incorrect subject–object directions. Therefore, these candidates are treated as inputs to the subsequent verification module.

To achieve precise relation alignment, we formulate relation classification as a natural language inference task. Specifically, the original sentence S is treated as the premise, and for each candidate entity pair (e_1, e_2) , a hypothesis space is constructed based on the predefined relation set R . Considering the uncertainty in relation directionality, we include both (e_1, r, e_2) and (e_2, r, e_1) , resulting in a hypothesis space of size $2|R|$.

Each premise–hypothesis pair is then fed into an NLI model to estimate the semantic consistency between the original sentence and the candidate hypothesis. Specifically, the NLI model outputs the prediction scores of different semantic categories, which are normalized into probabilities through a softmax function. The probabilities corresponding to the entailment and contradiction categories are denoted as $P_{\text{entailment}}$ and $P_{\text{contradiction}}$, respectively.

A differential scoring function is used to measure structural consistency:

$$S_{\text{score}} = P_{\text{entailment}} - P_{\text{contradiction}} \quad (3)$$

This score reflects the semantic alignment between the candidate relation and the original sentence. The hypothesis with the highest score is selected as the optimal alignment for the entity pair. To ensure reliability, a threshold constraint $S_{\text{score}} > \alpha$ is applied, retaining only high-confidence triplets. This process yields a set of reliable core triplets, providing a structural foundation for subsequent attribute verification.

3.6. Level 2: Attribute Consistency Verification

After obtaining the validated core triplets, we further extract and verify their qualified attributes. For each verified triplet $t = (e_1, r, e_2)$, a large language model is used again to generate candidate attribute values from the original sentence S . Since attributes serve as semantic modifiers of triplets and do not involve directionality, inverse relation modeling is not required.

To achieve accurate matching between attribute types and values, we construct an NLI-based verification mechanism. For each candidate attribute value v , we enumerate the predefined attribute type set A and construct formalized hypotheses; e.g., “the relation r between e_1 and e_2 has attribute a with value v .” These hypotheses are paired with the original sentence and fed into the NLI model to compute structural consistency scores.

Consistent with the triplet verification process, the hypothesis with the highest score is selected as the optimal match, and a threshold constraint $S_{\text{score}} > \alpha$ is applied. An attribute is attached to the triplet only if it achieves the global maximum score and satisfies the constraint. This process effectively avoids attribute mismatching and semantic drift, ensuring the integrity and consistency of the final hyper-relational structure.

To provide a clearer description of the overall workflow, the distribution-aware generation and hierarchical structural consistency verification processes are formally unified, as presented in Algorithm 1, where D_{LLM} denotes the set of synthetic samples generated by the large language model, and $D_{\text{candidate}}$ represents the candidate dataset combining the original and generated samples. R_{tail} denotes the set of long-tail relations identified according to the frequency threshold τ . H represents the hypothesis set constructed from

candidate relations or attributes, and S_{score} denotes the NLI-based consistency score. P_{entail} and $P_{\text{contradiction}}$ are the entailment and contradiction probabilities produced by the NLI model. The parameter α is the consistency verification threshold used to retain high-confidence triplets and attributes. v denotes a candidate attribute value generated during the attribute verification stage.

Algorithm 1 Overall Procedure of HRE-LSC

Require: Original dataset D , relation set R , attribute set A , threshold τ

Ensure: Enhanced dataset D^+

```

1:  $D_{\text{LLM}} \leftarrow \emptyset$ 
2: Stage 1: Distribution-aware Generation
3: Identify long-tail relation set  $R_{\text{tail}}$ 
4: for each  $r \in R_{\text{tail}}$  do
5:   Generate samples using LLM
6:   Add generated samples into  $D_{\text{LLM}}$ 
7: end for
8:  $D_{\text{candidate}} \leftarrow D \cup D_{\text{LLM}}$ 
9: Stage 2: Hierarchical Logical Verification
10:  $D^+ \leftarrow \emptyset$ 
11: for each sentence  $S \in D_{\text{candidate}}$  do
12:   Generate candidate entity pairs  $(e_1, e_2)$ 
13:   for each  $(e_1, e_2)$  do
14:     Construct hypothesis set  $H$ 
15:     Compute score:  $S_{\text{score}} = P_{\text{entail}} - P_{\text{contradiction}}$ 
16:     if  $S_{\text{score}} > \alpha$  then
17:       Generate attribute candidates
18:       for each  $v$  do
19:         Compute NLI score
20:       end for
21:       Select best attribute
22:       Add to result set
23:     end if
24:   end for
25: end for
26: return  $D^+$ 
  
```

3.7. Computational Complexity Analysis

The proposed hierarchical logical consistency verification strategy introduces additional computational costs due to the NLI-based consistency evaluation. Therefore, we analyze the computational complexity of the verification process.

In the Level 1 core triplet consistency verification stage, the relation hypothesis space is constructed for each candidate entity pair by considering both forward and inverse relation directions. The computational complexity of verifying a candidate entity pair is

$$O(|R| \times C_{\text{NLI}}) \quad (4)$$

where $|R|$ denotes the number of predefined relation types and C_{NLI} represents the computational cost of one NLI inference.

Assuming that the dataset contains N candidate samples and each sample contains an average of M candidate entity pairs, the overall complexity of Level 1 verification is

$$O(NM|R|C_{\text{NLI}}) \quad (5)$$

For the Level 2 attribute consistency verification stage, let N_t denote the number of validated core triplets after Level 1 verification, K represent the average number of

candidate attribute values for each triplet, and $|A|$ denote the size of the attribute set. The computational complexity of attribute verification is

$$O(N_t K |A| C_{NLI}) \quad (6)$$

Therefore, the overall computational complexity of the proposed hierarchical verification strategy is

$$O(NM|R|C_{NLI} + N_t K |A| C_{NLI}) \quad (7)$$

The above analysis indicates that the computational cost mainly comes from the NLI-based consistency verification process and increases linearly with the sizes of relation and attribute spaces. Therefore, the proposed method avoids exponential growth caused by enumerating relation combinations and maintains scalability when applied to HRE tasks with larger relation inventories.

4. Experiments and Results

4.1. Experimental Setup

4.1.1. Dataset

We conduct experiments on the standard sentence-level hyper-relational extraction dataset, HyperRED. This dataset contains rich complex factual structures, but exhibits a pronounced long-tail distribution at both the relation and attribute levels, along with pseudo-negative samples caused by unannotated attributes.

Based on the proposed HRE-LSC framework, we construct an augmented dataset, HyperRED+. Without altering the distribution of the original test set, augmentation is applied only to the training set. After distribution-aware generation and hierarchical structural consistency verification, the training set size increases from 48,598 triplets to 64,365, representing an improvement of approximately 32%. The number of entities grows from 130,170 to 175,987, as shown in Table 1.

Table 1. Statistics of the dataset before and after augmentation.

Dataset	Triplets	Entities
HyperRED	48,598	130,170
HyperRED+	64,365	175,987

For the distribution-aware data generation stage, we employ GPT-4o as the backbone large language model. Unless otherwise specified, the generation process uses temperature = 0.7 and greedy decoding with a maximum output length of 512 tokens. The same prompt template is applied across all generation experiments to ensure consistency and reproducibility.

To further validate the quality of the generated hyper-relational data, we conduct a human evaluation on randomly sampled subsets of D_{llm} . Specifically, we randomly select 200 generated hyper-relational facts three times independently, resulting in 600 evaluated instances in total.

Each instance is independently evaluated by two annotators with knowledge graph research experience. The annotators determine whether each generated hyper-relational fact is valid according to the corresponding context and knowledge structure. The evaluation considers three aspects: (1) semantic consistency, which measures whether the generated relation fact is consistent with the contextual semantics; (2) attribute validity, which evaluates whether the hyper-relation attributes and their values are appropriate for the corresponding relation; and (3) structural correctness, which verifies whether the generated samples follow the predefined hyper-relational format.

For conflicting annotations, a third annotator is invited to make the final decision. The evaluation scores are calculated based on the verified annotations. The final results are reported as the mean $\pm$ standard deviation over three independent sampling trials. The results are summarized in Table 2. The evaluation results demonstrate that the proposed distribution-aware generation strategy can effectively generate reliable hyper-relational data while maintaining semantic and structural consistency.

Table 2. Human evaluation results of generated hyper-relational data.

Evaluation Criterion	Accuracy (%)
Semantic Consistency	86.6 $\pm$ 1.0
Attribute Validity	78.4 $\pm$ 1.0
Structural Correctness	91.2 $\pm$ 0.6
Average	85.4 $\pm$ 0.8

4.1.2. Baselines

To comprehensively evaluate the effectiveness of the proposed approach, we consider three categories of representative baseline models:

- Text2NKG: A representative state-of-the-art (SOTA) method for hyper-relational extraction.
- CubeRE: An end-to-end extraction model based on cube-filling structural prediction.
- Generative Baseline: A BART-based model that formulates hyper-relational extraction as an end-to-end sequence generation task.
- Pipeline Baseline: A two-stage pipeline that first extracts relational triplets, and then utilizes BERT to extract qualifiers based on the identified triplets.
- GPT-3.5-turbo and GPT-4: Under zero-shot and few-shot settings, a unified prompt template is used to directly generate hyper-relational triplets, which are then parsed into structured outputs.

For fair comparison, “Ours (Text2NKG)” and “Ours (CubeRE)” denote the corresponding baseline models trained on the augmented HyperRED+ dataset generated by the proposed HRE-LSC framework, while maintaining the same backbone architecture and evaluation protocol as the original implementations.

4.1.3. NLI-Based Consistency Scoring

In the hierarchical structural consistency verification stage, we employ a pre-trained T5-XXL NLI model for inference without additional fine-tuning. For each candidate sample, a premise is constructed based on predefined templates (see Appendix A), and fed into the model to obtain the probabilities of entailment and contradiction. A differential score is then computed, and only samples with scores exceeding a predefined threshold are retained; otherwise, they are filtered out.

Unless otherwise stated, the NLI selection threshold α is set to 0.6 in all main experiments. Additional sensitivity analysis under different threshold settings is shown in Figure 3.

4.1.4. Evaluation Metrics

We adopt Precision, Recall, and F1-score as evaluation metrics.

During evaluation, a strict matching criterion is applied: a prediction is considered correct only if both the core triplet and all its associated qualified attributes are correctly identified.

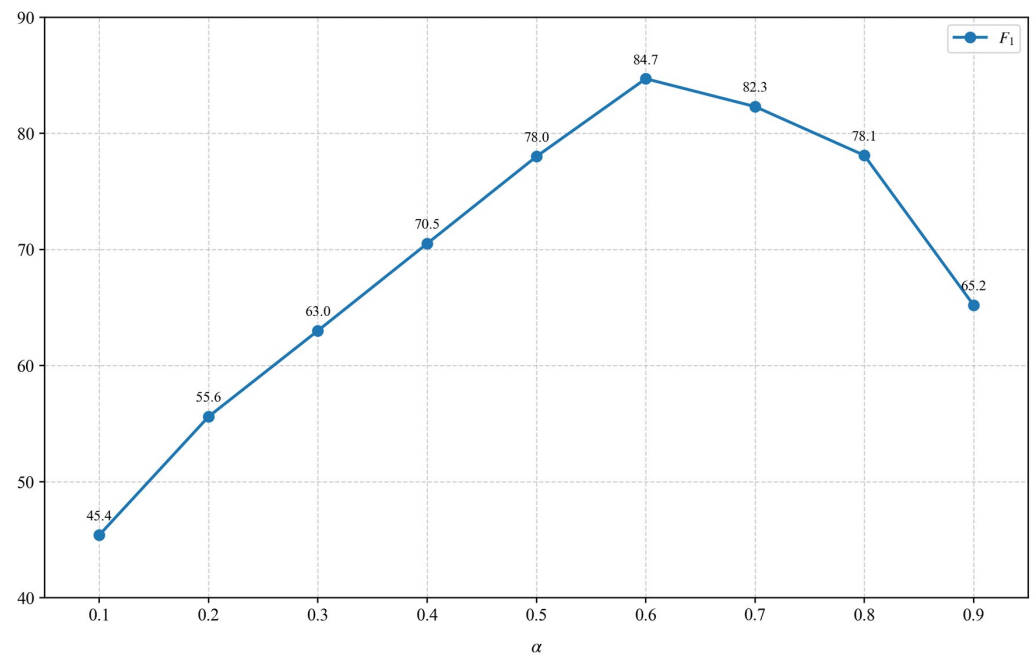


Figure 3. Sensitivity Analysis of NLI Selection Threshold.

4.2. Main Results

To evaluate the overall effectiveness of the proposed HRE-LSC framework, we conduct comparative experiments on the HyperRED test set. The results are reported in Table 3.

Table 3. Performance comparison on the HyperRED dataset.

Model	Precision	Recall	F1
GPT-4	0.158	0.154	0.156
GPT-3.5-turbo	0.114	0.109	0.111
Generative Baseline	64.60 $\pm$ 0.47	59.67 $\pm$ 0.35	62.03 $\pm$ 0.21
Pipeline Baseline	69.00 $\pm$ 0.48	57.55 $\pm$ 0.19	62.75 $\pm$ 0.29
CubeRE	65.82 $\pm$ 0.84	64.28 $\pm$ 0.25	65.04 $\pm$ 0.29
Ours (CubeRE)	68.70 $\pm$ 0.85	68.30 $\pm$ 0.31	68.50 $\pm$ 0.22
Text2NKG	91.06 $\pm$ 0.81	77.64 $\pm$ 0.46	83.81 $\pm$ 0.54
Ours (Text2NKG)	90.90 $\pm$ 0.80	79.30 $\pm$ 0.37	84.70 $\pm$ 0.46

From the experimental results, the following observations can be drawn.

4.2.1. Significant Overall Performance Improvement

Under the strict matching criterion, our method improves the F1 score of the Text2NKG baseline from 83.81% to 84.70%. Comparable improvements are also observed on the CubeRE model, where the F1 score increases from 65.04% to 68.50%. All results are reported as mean $\pm$ standard deviation over three independent runs with different random seeds. These results demonstrate that the proposed HRE-LSC framework effectively enhances the model's capability in capturing complex hyper-relational structures.

4.2.2. Limited Direct Extraction Capability of Large Language Models

Following Luo et al. [15], the reported results of GPT-3.5-turbo and GPT-4 are directly adopted from their study rather than reproduced in our experiments. These models achieve F1 scores of only 11.1% and 15.6%, respectively, which are substantially lower than those of structured extraction approaches. This comparison indicates that, without explicit

structural constraints and task-specific modeling, general-purpose large language models may struggle with high-precision hyper-relational extraction.

It is worth noting that the relatively low performance of large language models is not solely due to insufficient semantic understanding. Instead, it is largely influenced by the evaluation protocol of the HyperRED dataset, which adopts a predefined relation set with strict matching criteria. Even when generated outputs are semantically correct, they are considered incorrect if the predicted relations do not exactly match the predefined set, thereby significantly lowering the overall performance metrics.

4.3. Ablation Study

To analyze the contribution of each component, we conduct ablation experiments based on the Text2NKG model by removing key modules of HRE-LSC. The results are shown in Table 4.

Table 4. Ablation study of HRE-LSC.

Model	Precision	Recall	F1
Full Model	90.90 $\pm$ 0.80	79.30 $\pm$ 0.37	84.70 $\pm$ 0.46
w/o Distribution-aware Generation	89.41 $\pm$ 0.66	76.63 $\pm$ 0.28	82.54 $\pm$ 0.34
w/o Consistency Verification	74.22 $\pm$ 0.53	70.84 $\pm$ 0.24	72.58 $\pm$ 0.26
w/o Top-1 Strategy	73.50 $\pm$ 0.64	88.34 $\pm$ 0.35	80.29 $\pm$ 0.38
w/o Inverse Relation Modeling	90.07 $\pm$ 0.63	77.93 $\pm$ 0.21	83.84 $\pm$ 0.48

The ablation results reveal distinct impacts of different components. Removing the distribution-aware generation leads to decreases in both Precision and Recall, indicating that without alleviating the long-tail distribution, the model struggles to cover low-frequency relations and tends to overfit high-frequency ones, resulting in increased mismatches. Removing the NLI-based verification causes a substantial drop in F1 to 72.5%, along with significant declines in both Precision and Recall. This suggests that without structural consistency constraints, generated results are prone to relation boundary violations and attribute mismatches, which are penalized under strict evaluation criteria.

In contrast, removing the Top-1 selection strategy results in increased Recall but decreased Precision. This is because relying solely on thresholding retains multiple candidate relations, improving coverage but introducing semantic conflicts and redundant predictions. Removing inverse relation modeling leads to a slight decrease in F1 to 83.84%, indicating that this component mainly corrects relation direction errors.

Overall, NLI-based verification plays a central role in structural consistency modeling, the Top-1 strategy is crucial for conflict resolution, and distribution-aware generation improves coverage from the data perspective.

4.4. Further Analysis

4.4.1. Dataset Expansion Analysis

To illustrate the effectiveness of the distribution-aware augmentation strategy, we report the dataset scale before and after augmentation in Table 1.

The dataset size increases from 48,598 to 64,365, representing an approximate growth of 32%, with a substantial increase in the number of entities. To assess data quality, we conduct manual evaluation on sampled augmented data. The results show that the generated samples remain semantically consistent with the original text and do not introduce noticeable structural errors. This demonstrates that HRE-LSC effectively expands the dataset while maintaining data reliability.

4.4.2. Performance Analysis on Long-Tail Relations

To evaluate the effectiveness of the proposed method on long-tail relations, we conduct a fine-grained analysis on the low-frequency relation subset of the HyperRED dataset. The F1 performance of Text2NKG on HyperRED and HyperRED+ is illustrated in Figure 4.

Overall, HyperRED+ achieves more significant improvements on medium- and low-frequency relations, while the gains on high-frequency relations are limited or even exhibit slight fluctuations. This phenomenon reflects the different mechanisms constraining model performance across frequency regimes.

For high-frequency relations, performance changes are marginal. These relations typically exhibit stable and relatively simple semantic patterns, and the original dataset already provides sufficient coverage for effective learning. As a result, the marginal benefit of additional generated data is limited. Moreover, generated samples may introduce minor noise due to variations in expression granularity or contextual shifts, leading to slight performance fluctuations. This suggests that, for high-frequency relations, data quality is more critical than data scale.

In contrast, improvements on medium- and low-frequency relations are more pronounced. Fundamentally, these gains arise from the joint effect of distribution rebalancing and semantic coverage expansion. On one hand, the distribution-aware generation strategy increases the proportion of low-frequency relations, effectively mitigating the model's bias toward high-frequency ones. On the other hand, the generated data introduce diverse expression patterns, enhancing the model's ability to generalize across relation semantics.

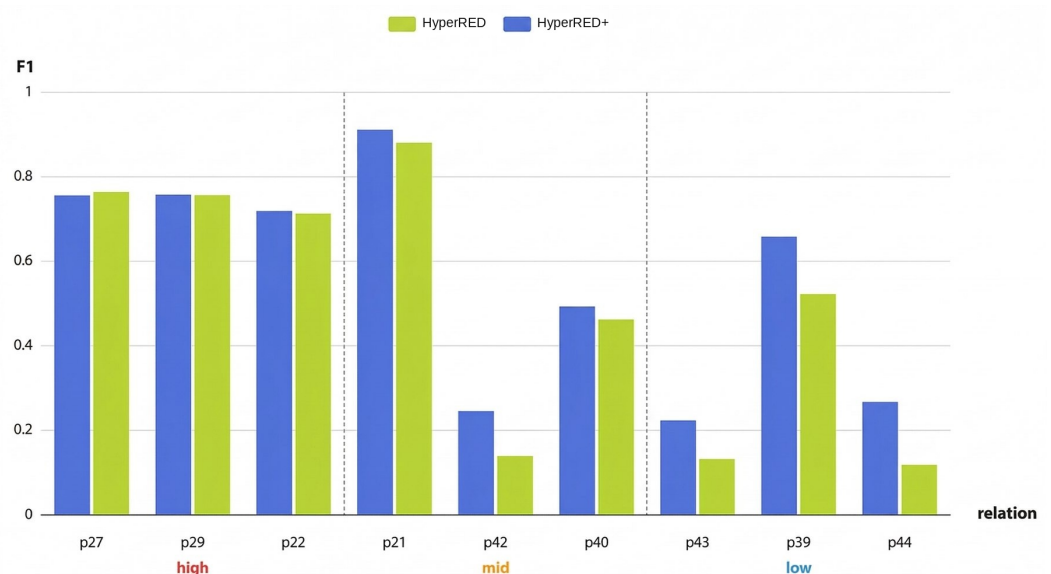


Figure 4. Comparison of F1 scores across different long-tail relation categories.

Furthermore, the improvement mechanisms differ between medium- and low-frequency relations. For medium-frequency relations, the original data already provide a moderate number of samples, but lack sufficient diversity; thus, generated data primarily contribute to semantic expansion and pattern completion. For low-frequency relations, where the model struggles to learn meaningful representations due to data scarcity, generated data mainly facilitate learning by rebalancing the distribution and providing essential structural patterns.

It is worth noting that the absolute performance on low-frequency relations still lags behind that of medium-frequency relations. This indicates that, although data generation alleviates data sparsity, complex structured relations still require stronger structural constraints to ensure data quality. Overall, these results validate the effectiveness of the

proposed method in long-tail scenarios and suggest that the primary performance bottleneck lies in data distribution bias and insufficient semantic coverage, rather than model capacity limitations.

4.4.3. Sensitivity Analysis of the Verification Threshold

To analyze the impact of the threshold parameter α in NLI-based verification, we conduct experiments under different settings. The results exhibit a typical inverted U-shaped trend, as shown in Figure 3.

The threshold α essentially controls the trade-off between precision and recall, functioning as a dynamic adjustment of the decision boundary under noisy distributions. When α is small, the filtering constraint is weak, allowing more noisy samples to be retained. Conversely, when α is too large, the constraint becomes overly strict, causing some valid samples to be incorrectly filtered out and reducing recall. Therefore, a moderate threshold achieves the optimal balance between noise suppression and information preservation.

4.4.4. Impact of Top- k Candidate Relations

To investigate the effect of the number of candidate relations on hyper-relational extraction performance, we conduct experiments with Top- k ranging from 1 to 5. The results are presented in Figure 5.

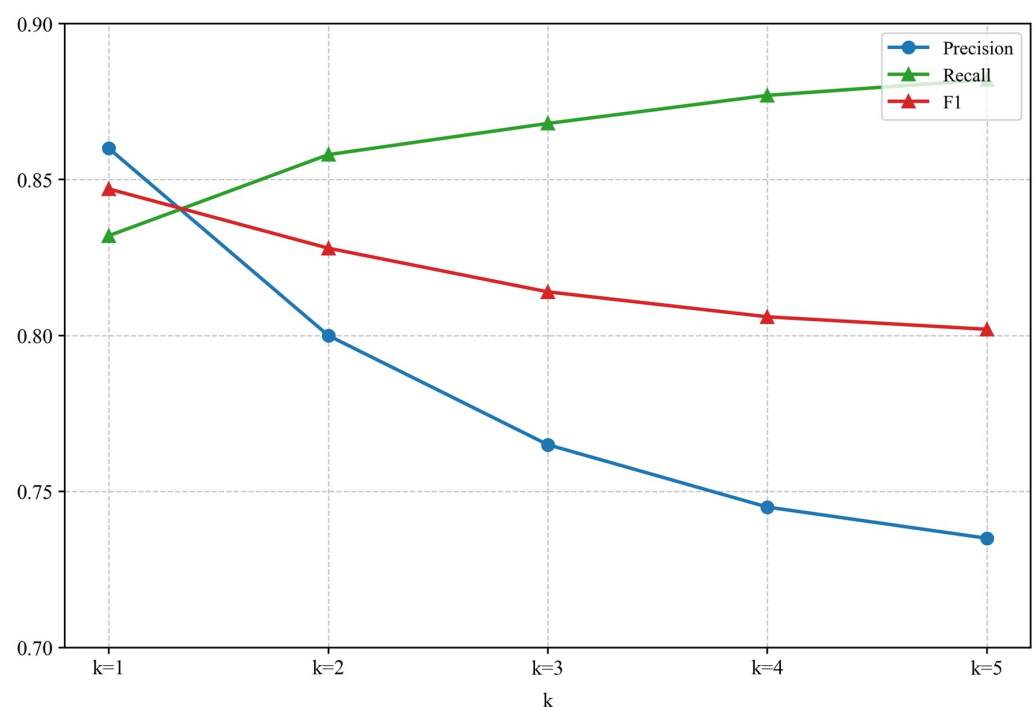


Figure 5. Impact of Top- k on hyper-relational extraction performance.

This phenomenon arises from the interaction between candidate expansion and threshold-based filtering. When k is small, newly introduced candidate relations are highly semantically relevant, and some high-confidence incorrect candidates may still pass the threshold, leading to a decline in precision. As k increases further, the overall confidence of additional candidates decreases, and many low-confidence candidates fail to pass the threshold. Consequently, the number of erroneous samples participating in final predictions grows only marginally, and the performance degradation trend becomes more gradual.

The results indicate that the threshold mechanism partially mitigates the noise introduced by increasing Top- k , but cannot fully replace a global optimal alignment strategy.

Overall, the combination of the Top-1 selection strategy and threshold constraint achieves the best balance between ensuring structural consistency and controlling noise.

5. Conclusions

In this paper, we propose HRE-LSC, a sentence-level hyper-relational data augmentation framework designed to improve the quality of training data for hyper-relational extraction. The proposed framework integrates distribution-aware synthetic data generation with hierarchical logical consistency verification, addressing the challenges of long-tail relation sparsity and unreliable samples generated by large language models.

The experimental results demonstrate that the proposed framework can effectively improve the performance of different HRE models, indicating that high-quality synthetic data with structural consistency can provide effective supervision for both generative and structured extraction paradigms. More importantly, this study highlights a general data-centric perspective for applying large language models in complex information extraction tasks. Instead of directly adopting LLM-generated outputs, the proposed framework treats generation as an intermediate step and introduces NLI-based semantic verification to identify and filter structurally inconsistent samples. This verification-driven paradigm provides a practical solution for improving the reliability of LLM-based data augmentation, especially in scenarios where explicit structural constraints are required.

Nevertheless, several limitations remain for future investigation:

- **Adaptive consistency threshold selection:** The current framework adopts a fixed consistency threshold α to filter unreliable generated samples. Although this strategy effectively improves data quality, the optimal threshold may vary across different datasets due to differences in relation distributions, noise levels, and structural complexity. Future work will investigate adaptive threshold selection strategies based on dataset characteristics to improve the robustness and generalization capability of the verification mechanism.
- **Lightweight verification models and generation cost reduction:** The current framework relies on large-scale NLI models for consistency verification and large language models for synthetic data generation, which may introduce considerable computational costs. Future work will explore lightweight NLI models and more efficient generation strategies to reduce computational overhead while maintaining verification accuracy and data quality.
- **Generalization across domains:** The current experiments are conducted mainly on the HyperRED benchmark. Future work should further evaluate the robustness of the proposed augmentation framework under cross-domain and low-resource settings to verify its applicability in broader scenarios.

Author Contributions: Methodology, X.D.; validation, Y.C.; writing—original draft, S.D.; writing—review & editing, S.X.; supervision, P.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the project “Digital Inheritance and Protection Platform Construction for the Cultural Heritage of the Chinese Eastern Railway and South Manchuria Railway”, grant number 26LYQ005L001.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: The authors would like to express their sincere gratitude to the anonymous reviewers for their insightful comments and constructive suggestions. The authors also thank their colleagues and research collaborators for valuable discussions and technical support during the course of this work.

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

Appendix A. Relation Construction Templates

To improve reproducibility and clarify the prompt design for relation construction, we summarize all relation-specific sentence templates used in our dataset construction.

Appendix A.1. Entity-to-Entity Relation Templates

- adjacent station: The station adjacent to sub. () is obj. ().
- award received: The award received by sub. () is obj. ().
- candidacy in election: sub. () was a candidate in the election obj. ().
- capital of: sub. () is the capital of obj. ().
- cast member: sub. () is a cast member of obj. ().
- chairperson: sub. () is the chairperson of obj. ().
- child: sub. () is a child of obj. ().
- coach of sports team: sub. () is the coach of the sports team obj. ().
- connecting line: The railway line connecting sub. () is obj. ().
- country: sub. () is located in the country obj. ().
- country of citizenship: The country of citizenship of sub. () is obj. ().
- director/manager: sub. () is the director or manager of obj. ().
- educated at: sub. () was educated at obj. ().
- employer: The person or organization for which sub. () works or worked is obj. ().
- followed by: sub. () was followed by obj. ().
- head of government: sub. () is the head of government of obj. ().
- head of state: sub. () is the head of state of obj. ().
- headquarters location: The headquarters of sub. () is located in obj. ().
- home venue: The home venue of sub. () is obj. ().
- instance of: sub. () is an instance of obj. ().
- league: sub. () participates in the league obj. ().
- location: The location of sub. () is obj. ().
- manufacturer: The manufacturer of sub. () is obj. ().
- member of: sub. () is a member of obj. ().
- member of political party: sub. () is a member of the political party obj. ().
- member of sports team: sub. () is a member of the sports team obj. ().
- occupation: The occupation of sub. () is obj. ().
- operator: The operator of sub. () is obj. ().
- owned by: sub. () is owned by obj. ().
- parent organization: The parent organization of sub. () is obj. ().
- part of: sub. () is part of obj. ().
- performer: sub. () is the performer of obj. ().
- place of birth: The place of birth of sub. () is obj. ().
- position held: sub. () held the position obj. ().
- residence: The residence of sub. () is obj. ().
- sport: sub. () plays the sport obj. ().
- spouse: sub. () is the spouse of obj. ().
- winner: sub. () is the winner of obj. ().

Appendix A.2. Event and Role-Based Templates

- participant: sub. () participated in obj. ().
- participating team: sub. () is a participating team in obj. ().
- nominated for: sub. () was nominated for obj. ().
- significant event: A significant event in the life or history of sub. () is obj. ().
- present in work: sub. () appears in the work obj. ().
- original broadcaster: The original broadcaster of obj. () is sub. ().
- voice actor: sub. () is the voice actor of obj. ().
- narrative role: sub. () has the narrative role obj. ().
- military branch: sub. () served in the military branch obj. ().
- replaces: sub. () replaced obj. ().
- followed by: sub. () was followed by obj. ().

Appendix A.3. Structural and Geographic Templates

- located in the administrative territorial entity: sub. () is located in the administrative territorial entity obj. ().
- located on street: sub. () is located on the street obj. ().
- shares border with: sub. () shares a border with obj. ().
- part of the series: sub. () is part of the series obj. ().
- subclass of: sub. () is a subclass of obj. ().
- used by: sub. () is used by obj. ().
- connecting line: The railway line connecting sub. () is obj. ().

Appendix A.4. Attribute-Level Construction Template

In addition to relation-specific templates, we define a generic attribute-based construction pattern:

Where the attribute “[ATTRIBUTE1]” has the value [VALUE1]

Appendix B. Long-Tail Relation Data Augmentation Prompt

Input:

- Target relation: {target_relation}
- Predefined hyper-relation attributes: {hyper_attribute_list}
- Minimum generation threshold: τ

Instruction: Generate new hyper-relational facts for the target relation to alleviate the long-tail distribution problem in the training dataset. The generated samples should preserve semantic consistency between entities, relations, and hyper-relation attributes.

Constraints:

- All generated samples must use only the given target relation {target_relation}.
- Do not introduce any new relation types.
- Hyper-relation attributes must be selected only from the predefined attribute list {hyper_attribute_list}.
- Each generated fact must contain at least one valid hyper-relation attribute.
- Attribute values should be semantically consistent with the corresponding subject, relation, and object.
- The generated entities must be realistic and consistent with the semantic meaning of the target relation.
- Avoid generating duplicate facts or trivial variations of existing samples.
- Do not hallucinate unsupported or contradictory facts.
- Generate no fewer than τ hyper-relational facts for the target relation.

Output format: (subject, relation, object, {(attribute_1, value_1), (attribute_2, value_2), ...})

Appendix C. Prompt Templates for Relation and Hyper-Relation Extraction

We design a two-stage prompting strategy to guide structured relation extraction and hyper-relational construction.

Appendix C.1. Stage-1: Relation Triple Extraction Prompt

Input:

- Sentence: {sentence}
- Relation set: {relation_list}

Instruction: Extract all valid (subject, relation, object) triples from the sentence using only relations from the given set. Do not introduce any new relations.

Constraints:

- Only use relations in the provided set.
- Extract at least 15 triples if possible.
- Do not hallucinate or fabricate facts.
- The subject and object entities must be exact spans from the original sentence (no paraphrasing, no normalization, no entity rewriting).

Output format: (subject, relation, object)

Appendix C.2. Stage-2: Hyper-Relational Construction Prompt

Input:

- Sentence: {sentence}
- Trusted triples: {triples}
- Attribute list: {attribute_list}

Instruction: Extend each trusted triple by selecting valid attributes from the attribute list to construct hyper-relational triples.

Constraints:

- Only use attributes in the provided list.
- Each output must follow the hyper-relational structure.
- Attribute values must be grounded in the sentence or trusted triples.
- No hallucination or fabricated content.
- The subject and object entities must be exact spans from the original sentence (no paraphrasing, no normalization, no entity rewriting).
- Generate at least 15 hyper-relational triples.

Output format: (subject, relation, object, (attribute, attribute_value))

Appendix D. Attribute Index Mapping

In this section, we define the mapping between attribute identifiers and their corresponding semantic meanings. These predefined attributes are used in the hyper-relational construction process described in this work.

- p1: instance of
- p2: has part
- p3: diocese
- p4: academic degree
- p5: sports league level
- p6: academic major

- p7: position played on team/speciality
- p8: street number
- p9: object has role
- p10: number of matches played / races / starts
- p11: electoral district
- p12: ticker symbol
- p13: affiliation
- p14: country
- p15: number of points / goals / set scored
- p16: national team appearances
- p17: mother
- p18: ranking
- p19: series ordinal
- p20: position held
- p21: publication date
- p22: character role
- p23: located in the administrative territorial entity
- p24: statement is subject of
- p25: point in time
- p26: together with
- p27: start time
- p28: applies to part
- p29: end time
- p30: subject has role
- p31: member of political party
- p32: of
- p33: location
- p34: quantity
- p35: replaces
- p36: connecting line
- p37: follows
- p38: for work
- p39: performer
- p40: statement disputed by
- p41: towards
- p42: nominee
- p43: winner
- p44: adjacent station

References

1. Ji, S.; Pan, S.; Cambria, E.; Marttinen, P.; Yu, P.S. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 494–514. [[CrossRef](#)]
2. Dai, Y.; Wang, S.; Xiong, N.N.; Guo, W. A survey on knowledge graph embedding: Approaches, applications and benchmarks. *Electronics* **2020**, *9*, 750. [[CrossRef](#)]
3. Zhang, J.; Zhang, H.; Zhang, D.; Yong, L.; Huang, S. End-to-end beam retrieval for multi-hop question answering. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 1718–1731.
4. Chen, R.; Jiang, W.; Qin, C.; Rawal, I.S.; Tan, C.; Choi, D.; Xiong, B.; Ai, B. LLM-Based Multi-Hop Question Answering with Knowledge Graph Integration in Evolving Environments. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024*; Al-Onaizan, Y., Bansal, M., Chen, Y.N., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 14438–14451.

5. Bahrami, R.; Yahyapour, R. Re-Representation in Sentential Relation Extraction with Sequence Routing Algorithm. In *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 315–327.
6. Yan, T.; Zhang, X.; Luo, Z. LTACL: Long-tail awareness contrastive learning for distantly supervised relation extraction. *Complex Intell. Syst.* **2024**, *10*, 1551–1563.
7. Li, Y.; Zhou, K.; Qiao, Q.; Wang, Q.; Li, Q. Re-examine distantly supervised NER: A new benchmark and a simple approach. In *Proceedings of the 31st International Conference on Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 10940–10959.
8. Xie, C.; Liang, J.; Liu, J.; Huang, C.; Huang, W.; Xiao, Y. Revisiting the negative data of distantly supervised relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 3572–3581.
9. Zeng, D.; Zhu, J.; Chen, H.; Dai, J.; Jiang, L. Document-level denoising relation extraction with false-negative mining and reinforced positive-class knowledge distillation. *Inf. Process. Manag.* **2024**, *61*, 103533. [\[CrossRef\]](#)
10. Swarup, A.; Pan, T.; Wilson, R.; Bhandarkar, A.; Woodard, D. LLM4RE: A data-centric feasibility study for relation extraction. In *Proceedings of the 31st International Conference on Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 6670–6691.
11. Hu, X.; Liu, A.; Tan, Z.; Zhang, X.; Zhang, C.; King, I.; Yu, P.S. GDA: Generative data augmentation techniques for relation extraction tasks. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 10221–10234.
12. Li, J.; Luo, X.; Lu, G.; Zhang, S. Dhrl4hkg: A dual-hypergraph representation learning for hyper-relational knowledge graphs. *Knowl.-Based Syst.* **2025**, *325*, 113886. [\[CrossRef\]](#)
13. Chia, Y.K.; Bing, L.; Aljunied, S.M.; Si, L.; Poria, S. A dataset for hyper-relational extraction and a cube-filling approach. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 10114–10133.
14. Saini, P.; Sarkar, C.; Nayak, T. Are Relational Triple Extraction Frameworks Sufficient for Hyper-relational Facts? In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 112–119.
15. Luo, H.; Yang, Y.; Yao, T.; Guo, Y.; Tang, Z.; Zhang, W.; Peng, S.; Wan, K.; Song, M.; Lin, W.; et al. Text2nkg: Fine-grained n-ary relation extraction for n-ary relational knowledge graph construction. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 27417–27439. [\[CrossRef\]](#)
16. Yao, Y.; Ye, D.; Li, P.; Han, X.; Lin, Y.; Liu, Z.; Liu, Z.; Huang, L.; Zhou, J.; Sun, M. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 764–777.
17. Wei, T.; Shi, J.X.; Zhang, M.L.; Li, Y.F. Robust long-tailed learning under label noise. *Front. Comput. Sci.* **2025**, *20*, 2001321. [\[CrossRef\]](#)
18. Zhang, C.; Alpanidis, G.; Fan, G.; Deng, B.; Zhang, Y.; Liu, J.; Kamel, A.; Soda, P.; Gama, J. A systematic review on long-tailed learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 13670–13690. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Qiu, J.; Lin, Y.; Chen, H.; Cheng, D.; Shafiq, M.; Zhang, L.; Tian, Z. Diverse classifiers with label dependencies for long-tail relation extraction in big data. *Comput. Electr. Eng.* **2023**, *110*, 108812. [\[CrossRef\]](#)
20. Tang, J.; Li, L.; Lu, H.; Zhang, B.; Wu, H. Improving long-tail relation extraction via adaptive adjustment and causal inference. *Neurocomputing* **2023**, *552*, 126563. [\[CrossRef\]](#)
21. Meng, S.; Hu, X.; Liu, A.; Li, S.; Ma, F.; Yang, Y.; Wen, L. RAPL: A Relation-Aware Prototype Learning Approach for Few-Shot Document-Level Relation Extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*; Bouamor, H., Pino, J., Bali, K., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 5208–5226.
22. Menotti, L.; Marchesin, S.; Silvello, G. DOREMI: Optimizing Long Tail Predictions in Document-Level Relation Extraction. *Knowl.-Based Syst.* **2026**, *337*, 115359. [\[CrossRef\]](#)
23. Meng, Z.; Zhang, W.; Zhao, M.; Xu, T.; Yu, J.; Yu, R.; Yu, M. DSPrompt: Prompt learning with relation abstraction and context injection for distant supervised relation extraction. In *ECAI 2024*; IOS Press: Amsterdam, The Netherlands, 2024; pp. 3915–3922.
24. Tan, Q.; Xu, L.; Bing, L.; Ng, H.T.; Aljunied, S.M. Revisiting DocRED-addressing the false negative problem in relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 8472–8487.
25. Hao, K.; Yu, B.; Hu, W. Knowing false negatives: An adversarial training method for distantly supervised relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 9661–9672.

26. Chen, J.W.; Fu, T.J.; Lee, C.K.; Ma, W.Y. H-FND: Hierarchical false-negative denoising for distant supervision relation extraction. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 2579–2593.
27. Chen, X.; Wu, H.; Shi, X. Consistent Prototype Learning for Few-Shot Continual Relation Extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Rogers, A., Boyd-Graber, J., Okazaki, N., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 7409–7422.
28. Feng, S.Y.; Gangal, V.; Wei, J.; Chandar, S.; Vosoughi, S.; Mitamura, T.; Hovy, E. A survey of data augmentation approaches for NLP. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 968–988.
29. Diaz-Garcia, J.A.; Lopez, J.A.D. A survey on cutting-edge relation extraction techniques based on language models. *Artif. Intell. Rev.* **2025**, *58*, 287. [[CrossRef](#)]
30. Li, J.; Jia, Z.; Zheng, Z. Semi-automatic data enhancement for document-level relation extraction with distant supervision from large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 5495–5505.
31. Zheng, Y.; Ke, W.; Liu, Q.; Yang, Y.; Zhao, R.; Feng, D.; Zhang, J.; Fang, Z. Making LLMs as Fine-Grained Relation Extraction Data Augmentor. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*; Larson, K., Ed.; International Joint Conferences on Artificial Intelligence Organization: Montpellier, France, 2024; pp. 6660–6668.
32. Popovic, N.; Kangen, A.; Schopf, T.; Färber, M. DocIE@ XLLM25: In-Context Learning for Information Extraction using Fully Synthetic Demonstrations. In *Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 298–309.
33. Josifoski, M.; Sakota, M.; Peyrard, M.; West, R. Exploiting asymmetry for synthetic training data generation: SynthIE and the case of information extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 1555–1574.

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.