ODDA: An OODA-Driven Diverse Data Augmentation Framework for Low-Resource Relation Extraction

Yijie Zhong[1], **Yunfan Gao**[2], **Xiaolian Zhang**[3], **Haofen Wang**[1]

[1] College of Design and Innovation, Tongji University, China,
[2] Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University, China,
[3] Huawei Technologies Co. Ltd., China,

Email: dun.haski@gmail.com, carter.whfcarter@gmail.com

Abstract

Data Augmentation (DA) has emerged as a promising solution to address the scarcity of high-quality annotated data in low-resource relation extraction (LRE). Leveraging large language models (LLMs), DA has significantly improved the performance of RE models with considerably fewer parameters. However, existing DA methods struggle with diversity misalignments, as they neglect the diversity required by the model and generate homogeneous augmentations that do not cover the inter-sample and inter-relation variability, leading to suboptimal performance. Inspired by the Observe-Orient-Decide-Act (OODA) framework, which provides a robust theoretical foundation for iterative decision-making under dynamic conditions, we propose an OODA-driven Diverse DA method (ODDA), guiding the data generation and selection process. ODDA first observes the RE model's behavior to select effective demonstrations for LLMs. Next, it orients LLMs towards generating diverse data by replacing schema constraints with attribute constraints. Then ODDA decides on the final augmented dataset with overall diversity from a global search and finally acts to train the RE model. Extensive experiments on three widely-used benchmarks demonstrate that ODDA consistently outperforms state-of-the-art baselines, achieving average F1 improvements of 3.1% across various LRE scenarios while maintaining enhanced model stability.

1 Introduction

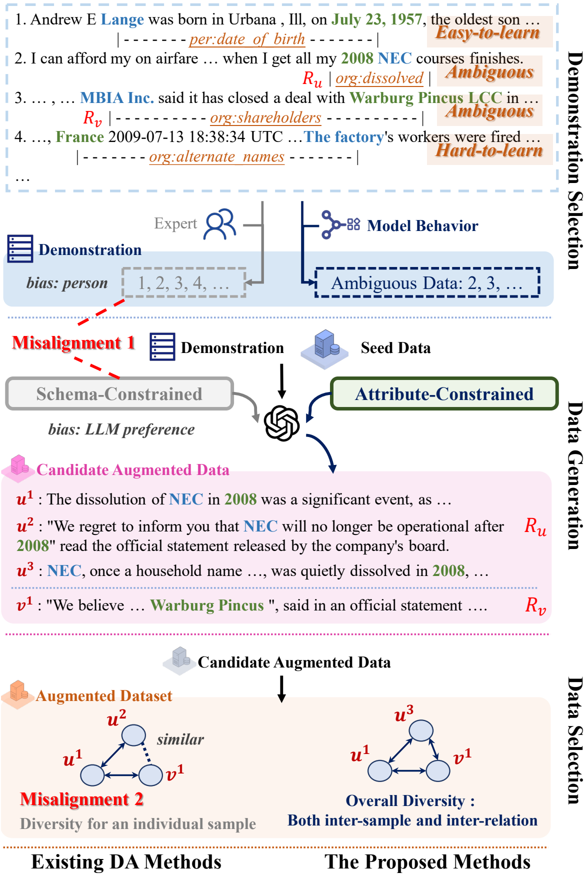
Relation Extraction (RE) aims to identify semantic relations between given entity pairs and converts unstructured text into structured triplets. It supports critical applications like knowledge graph construction (Zhong et al., 2025) and intelligent question answering (Molfese et al., 2024). Training well-designed RE models with a supervised paradigm is currently the mainstream and effective strategy. However, obtaining large-scale high-quality annotated data is laborious and expensive, which limits the applicability of supervised RE systems in low-resource scenarios (LRE) (Deng et al., 2024).

Data Augmentation (DA) offers a direct and efficient approach for LRE compared to meta-learning (Veyseh et al., 2023) or transfer learning (Gururaja et al., 2023). Despite Large Language Models (LLMs) possessing zero/few-shot capabilities, they still struggle on RE tasks (Li et al., 2023a; Han et al., 2023) with rich pre-defined patterns, complex classification spaces (Xu et al., 2023b), and high computational costs. By leveraging LLMs' ability to perform the inverse task of generating sentences from triples (Ma et al., 2024), DA forms a bridge between LLMs and RE models, ultimately boosting LRE performance.

The effectiveness of DA heavily depends on its quality and diversity, which plays a pivotal role in RE performance (Yu et al., 2023), particularly in capturing complex relation patterns and semantic variations. This paper addresses the critical challenge of generating highly diverse and semantically valid data for LRE tasks. Traditional DA methods (Cai et al., 2020; Min et al., 2020), relying on surface-level operations like token replacement ad deletion, often yield suboptimal data quality and limited diversity. GDA (Hu et al., 2023) advances this field by introducing generative models and establishing diversity as a crucial consideration in DA. Subsequent work enhances data diversity through structured guidance, incorporating schema (Xu et al., 2023b) and keywords (Zheng et al., 2024) into LLM-based generation.

Generally, the LLM-based DA methods consist of three parts: demonstration selection which selects examples to guide the LLMs, data generation which activates the LLMs to generate candidate sentences and data selection which picks candidate data to build the final augmented dataset. As shown in Figure 1, they suffer from two key misalignments

Findings of the Association for Computational Linguistics: ACL 2025, pages 267–285
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics



1. Andrew E Lange was born in Urbana , Ill, on July 23, 1957, the oldest son …
|- - - - - - - per:date_of_birth - - - - - - - |

2. I can afford my on airfare … when I get all my 2008 NEC courses finishes.
R_u | org:dissolved |

3. … , … MBIA Inc. said it has closed a deal with Warburg Pincus LCC in …
R_v |- - - - - - - - - - - org:shareholders - - - - - - - - - - - |

4. …, France 2009-07-13 18:38:34 UTC …The factory's workers were fired …
|- - - - - - - org:alternate_names - - - - - - - |

…

u^1 : The dissolution of NEC in 2008 was a significant event, as …

u^2 : "We regret to inform you that NEC will no longer be operational after 2008" read the official statement released by the company's board.

u^3 : NEC, once a household name …, was quietly dissolved in 2008, …

R_u

v^1 : "We believe … Warburg Pincus ", said in an official statement ….

R_v

Figure 1: Comparison of existing DA methods, which have two diversity misalignments, and the proposed method based on the OODA framework.

that hinder the diversity of the augmented dataset.
Misalignment 1: The diversity injected by the method design does not match the diverse data needed by the target RE model. The human-driven selection for demonstrations and schema-constrained prompting during generation introduce LLMs' inherent preferences and prevent producing truly varied candidate augmented data.
Misalignment 2: Focusing solely on individual sample diversity leads to what we term as '*homogeneous diversity*', thereby neglecting the overall inter-sample and inter-relation variability.

The misalignments motivate our solution driven by the Observe-Orient-Decide-Act (OODA) theory, which asserts that both generation and selection processes must be dynamically guided by the RE model's learning behavior and overall situation. Therefore, we introduce an OODA-driven Diverse DA method, called ODDA, that enhances the diversity and quality of augmented datasets for LRE. ODDA adapts the decision-making paradigm by observing the target model's behavior, orienting the data generation process via attribute constraints, deciding upon the most diverse and effective samples, and ultimately acting by training the RE model. Specifically, we propose three key components: 1) Selective Demonstration Filtering that identifies samples with a moderate learning difficulty to guide LLMs effectively; 2) Attributed-Constrained Data Generation that incorporates diverse linguistic variations (*e.g.* syntactic structures, semantic patterns); and 3) Overall Diversity Data Selection that optimizes both inter-sample and inter-relation diversity through global optimization. ODDA also enables continuous optimization of the augmented data and the RE model by restarting the whole OODA loop. We carefully analyze the diversity and quality of the augmented data and experiments validate that ODDA produces more diverse data and achieves superior performance in LRE scenarios. The contributions of this paper are as follows:

- We identify two critical misalignments in existing DA methods for LRE tasks that limit data diversity. Then we propose the first diverse DA method guided by OODA-theory.

- We design a demonstration selection mechanism based on model behavior coupled with attribute-constrained prompting to generate data the RE model truly needs.

- We select the candidate augmented data under an overall perspective that enhances inter-sample and inter-relation variability.

- Comprehensive experiments on three widely used benchmarks show that ODDA consistently achieves state-of-the-art and stable performance (∼4% F1↑ in 8-shot settings) with diverse augmented samples.

2 Related Work

Various approaches are explored in recent years to address LRE challenges (Gao et al., 2025; Oida-Onesa and Ballera, 2024): meta-learning (Hu et al., 2021; Liu et al., 2022b; Veyseh et al., 2023), transfer learning (Sarhan and Spruit, 2020; Gururaja et al., 2023), instruction prompting (Li et al., 2023b), and data augmentation (Zhang et al., 2024). For DA, traditional methods augment data by substituting tokens, like synonym replacement (Mueller and Thyagarajan, 2016) or token-level operations like random insertion, swap, and deletion (Wei and Zou, 2019), which is further

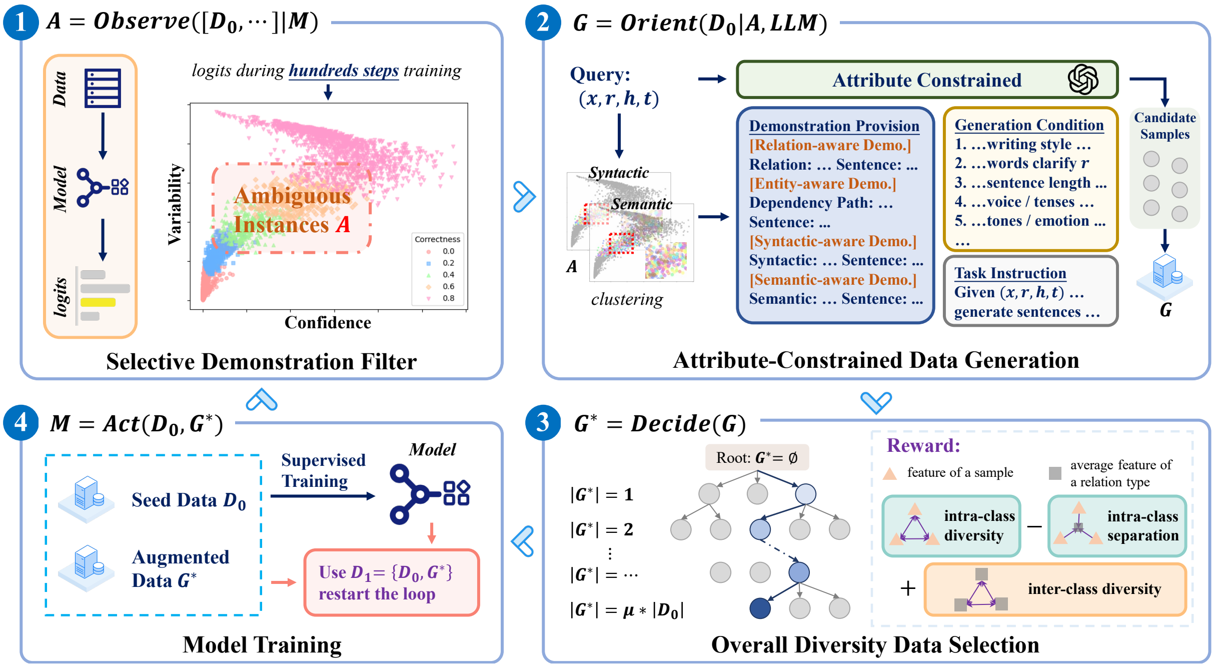


Figure 2: The proposed ODDA framework for RE and we implement the OODA loop through: 1) Model-guided observation for demonstration filtering, extending beyond the existing methods' selection $A = \texttt{Observe}([D_o, \cdots])$; 2) Attribute-constrained generation with filtered demonstrations and generation conditions; 3) Diversity-optimized augmented instance selection with global search while existing methods only consider $G^\star = \bigcup_{g \in G} \texttt{Decide(g)}$; 4) Model training with iterative refinement capability. Each stage contributes to addressing the diversity misalignments.

enhanced by leveraging word embeddings to generate contextually similar replacement (Jiao et al., 2020). While these methods lack diversity, later work like Back-translation (Fabbri et al., 2021), masked language modeling (Lowell et al., 2021), and GPT-2 fine-tuning (Anaby-Tavor et al., 2020) utilize generative models but struggle with relation consistency (Chen, 2024). Recent research, including GDA (Hu et al., 2023), ConsistRE (Zheng et al., 2024), and UnleashLLMRE (Xu et al., 2023b), employ LLMs to generate data based on the given triplets. They utilize schema constraints to guide LLMs along with an arbitrary selection of data, resulting in a lack of data diversity.

3 Preliminary

Given an original dataset D consisting of n instances $\{x_1, \cdots, x_n\}$, each instance x is annotated with a triplet (r, h, t), representing the relation type, head entity, and tail entity, respectively. For any relation type c, we denote the subset of instances with this relation as $R_c = \{x | r = c\}$. In low-resource scenarios, the $k-$shot setting restricts the size of each R_c to at most k. The goal of the data augmentation (DA) is to generate μ new instances x' for each origin instance x while keeping the annotations (r, h, t) unchanged. Ultimately, all instances x and their augmented data x' are then used to train a relation extraction (RE) model M.

4 Methodology

4.1 Overview

Figure 2 shows the workflow of ODDA with the OODA framework. ODDA first selects demonstrations by observing the RE model's behavior. Next, it achieves candidate data from attribute-constrained generation. Then, the final diverse augmented dataset is decided from a global search. Finally, the augmented and seed data are used to train the RE model. We could initialize a new loop with the augmented data and model after training.

4.2 Selective Demonstration Filtering

`Observe` stage: A subset of seed data most effective for the RE model's learning is identified as demonstrations from the behavior during the model training.

The goal of this stage is to select demonstrations that are beneficial to the model's learning process. Previous methods depend on experts' subjective interpretations of diversity to guide data generation, without confirming the data effectively optimizes the target RE model. Such diversity misaligns with the RE model's actual needs, undermining the benefits of DA. Inspired by the *Data Map* (Swayamdipta et al., 2020) that dynamically characterizes and diagnoses a dataset during training, offering fresh insights for dataset optimization, we propose a selective demonstration filtering (SDF) strategy. SDF leverages the target model's behavior to avoid generating data that could impair learning.

Specifically, given an available dataset D_0, we train the target model M for a few hundred steps - *typically just a few minutes* - and record the *logits* of labeled relation type for each instance at set intervals. By analyzing the mean $confidence$ and variance $variability$ of the *logits* throughout training, we assign each instance x a value $V(x) = (var_x, con_x)$ to create the *Data Map*. This allows us to categorize instances as easy-to-learn, ambiguous, or hard-to-learn. Since M shows little improvement on both easy-to-learn and hard-to-learn instances, we filter these out to produce an ambiguous data subset $A = \{x|(\tau, \tau) \leq V(x) \leq (\epsilon, \epsilon)\}$, where τ and ϵ are thresholds. We then generate more effective data by referencing A in the ICL.

4.3 Attribute-Constrained Data Generation

> `Orient` stage: LLMs generate candidate dataset from seed dataset, guided by attribute-constrained provided demonstrations and generation conditions.

The goal of this stage is to generate diverse augmented data for each seed data while reducing person's and LLMs' biases. Relying solely on class-conditional prompts or schema-constrained generation (Xu et al., 2023b) (*i.e.* offering relation type or expected triplets) as guidelines for LLMs can introduce LLMs' inherent biases (Yu et al., 2023), limiting data diversity. Thus, we propose attribute-constrained generation, which guides LLMs using multiple independent attributes that influence both *demonstration* and *generation conditions*.

For demonstration provision, we select demonstrations from four attributes for each seed instance (x, r, h, t): *relation*, *entity*, *semantic*, and *syntax*. **For relation-aware demonstrations**, we choose two random instances from the ambiguous subset A that share the same relation type r to form E^t_x. **For entity-aware demonstrations**, we first calculate the dependency path between the head and tail entity of each instance in A and instance x. Then we select three instances with the shortest, longest, and identical dependency path length to x from A to create E^e_x. The dependency path and the corresponding tokens are highlighted in the prompt. **For the semantic and syntax demonstrations**, we extract the semantic[1] and syntactic[2] features for x and each instance in A. To enhance instance coverage and diversity, we cluster the instances in A and select one instance from both the closest and furthest clusters to x, forming E^s_x and E^l_x respectively. The prompt instructs LLMs to emphasize the semantics or syntactic structures outlined in these demonstrations. All selected demonstrations $\{E^t_x, E^e_x, E^s_x, E^l_x\}$ are provided to the final prompt.

For generation conditions, we not only preserve the consistency of (r, h, t) but also integrate several attributes as constraints. To ensure these attributes are both rational and comprehensive, we employ a human-ai collaboration scheme (Liu et al., 2022a; Wiegreffe et al., 2022) to select the most suitable, high-quality attributes for the RE task. We prompt the LLM with questions such as "*What do you think are important attributes to generate diverse sentences under ...?*" to identify the best candidates. Using attribute-constrained demonstrations and generation conditions, along with schema instruction, LLMs generate a highly enriched candidate augmented dataset $G^\star$ (More details and the prompt format are provided in Appendix C).

4.4 Overall Diversity Data Selection

> `Decide` stage: Select data from the candidate augmented dataset with an overall perspective to construct the final diverse augmented dataset.

The goal of this stage is to build a diverse augmented dataset. Data selection is crucial for DA, yet existing methods primarily filter an individual seed sample's augmented data without addressing overall augmented dataset diversity. Selecting homogeneous samples from different seed samples reduces the dataset diversity. Thus, we propose an overall diversity data selection approach that considers both inter-sample and inter-relation diversity.

We apply a global search strategy and leverage *Monte Carlo Tree Search* (MCTS) to explore various sample combinations and optimize for diversity. MCTS evaluates multiple possibilities and guides the selection toward maximizing dataset diversity.

[1] The semantic features are obtained from sentence-BERT or Embedding APIs (such as `text-embedding-3-small`).

[2] The syntactic features are constructed by the depth of the syntax tree, the size of subtrees, the distribution of dependency paths among sentence tokens, the distribution of dependency types, sentence length, and the count of punctuation marks.

Accordingly, we design a specialized reward function to represent the overall diversity.

Let the features of an augmented sample be z_i (for both semantic and syntactic features). To ensure sufficient **intra-class diversity** for relation type c, the differences among the samples must be substantial, which can be formulated as:

$$\mathcal{R}_{div}^{intra}(c) = \frac{2}{|R_c||R_c - 1|} \sum_{(i,j)\in R_c} d(z_i, z_j), \quad (1)$$

where $d(\cdot)$ calculates the distance between two features. Since samples from different relation types should maintain discernibility, we introduce a penalty term for **intra-class separation** to prevent excessive dispersion among them:

$$\mathcal{R}_{sep}^{intra}(c) = \frac{1}{|R_c|} \sum_{i\in R_c} d(z_i, \mu_c), \quad (2)$$

where μ_c represents the average features of all samples belonging to the relation type c. To promote greater differences between distinct relation types, we encourage larger **inter-class diversity**:

$$\mathcal{R}_{div}^{inter} = \underset{c\neq l}{min}\, d(\mu_c, \mu_l). \quad (3)$$

Combining these considerations, we arrive at our reward function $\mathcal{R}_{total} = \exp(\mathcal{R}_{div}^{inter} + \mathcal{R}_{div}^{intra} - \mathcal{R}_{sep}^{intra})$. Finally, we achieve the augmented dataset $G^\star$ that preserves overall diversity, ensuring better performance for the RE model.

4.5 Model Training

> `Act` stage: The diverse augmented dataset and seed dataset are combined to train a reliable and stable M.

We combine the selected augmented dataset $G^\star$ with the $k-$shot dataset D_0 to form the final training dataset $D^\star = G^\star \cup D_0$, which is used to train the RE model M. Notably, the proposed method allows for continuous optimization of both the augmented data and the model. Thus, $D^\star$ can serve as the new initial dataset to restart the OODA loop.

5 Experiments

5.1 Experimental Settings

Datasets. We conduct experiments on three public relation extraction datasets: TACRED (Zhang et al., 2017), Re-TACRED (Stoica et al., 2021), and SemEval (Hendrickx et al., 2019). The statistics of the datasets are presented in Table 1. More details about the datasets can be found in Appendix A.

Dataset	# Rel	# Train				# Val	# Test
		8-shot	16-shot	48-shot	All		
TACRED	42	334	662	1954	68124	22631	15509
ReTACRED	40	318	630	1858	58465	19584	13418
SemEval	19	144	288	860	6507	1493	2717

Table 1: Numbers of instances in our experimental datasets. *k-shot* denotes sampling *n* instances from each relation type. When a relation type has fewer than *k* instances, we sample all available data. *All* refers to the complete training dataset.

In this study, we sample 8, 16, and 48 instances for each relation type to simulate low-resource scenarios, following (Xu et al., 2023b; Zheng et al., 2024). All methods generate augmented data with an 8× expansion on the sampled instances. Each result is reported over five runs with different seeds.

Metrics. For evaluating RE performance, we adopt the widely used *Micro-F1*. To assess the diversity of the generated samples, we evaluate from both lexical and semantic perspectives. For lexical diversity, we employ *Type-Token Ratio* (TTR) (Tweedie and Baayen, 1998), *Distinct-N* (Li et al., 2016), and *Self-BLEU* (Zhu et al., 2018), while semantic diversity is measured using *Average Pairwise sample Similarity* (APS) (Mishra et al., 2020) by computing inter-class and intra-class APS scores. Beyond sufficient diversity, we also evaluate sample quality using *MAUVE** and *Front-Integral** (Pillutla et al., 2023). More details can be found in Appendix B.

Compared Methods. We choose 5 DA methods for comparison. (1) WordNet Synonym Substitution (WSS) (Mueller and Thyagarajan, 2016): replacing tokens with synonyms from WordNet; (2) Word Embedding Substitution (WES) (Jiao et al., 2020): replacing tokens with contextual embeddings from BERT; (3) LAMBADA (Anaby-Tavor et al., 2020): fine-tuning generative models to generate candidate examples. Additionally, we include LLM-based DA methods: (4) UnleashLLMRE (Xu et al., 2023b): introducing data generation with LLM to boost previous RE solutions; (5) ConsistRE (Zheng et al., 2024): utilizing LLM to generate consistent and triplet-preserving samples.

As DA methods are data-centric, we keep the RE baseline models simple yet competitive (Xu et al., 2023a). In line with previous work, we select two RE models to ensure a fair comparison of each DA method. (1) TYPMarker (Zhou and Chen, 2022): a fine-tuning method that leverages entity typed markers; (2) KnowPrompt (Chen et al., 2022): a prompt-tuning method that uses knowledge-aware

	Method	TACRED			ReTACRED			SemEval		
		8-shot	16-shot	48-shot	8-shot	16-shot	48-shot	8-shot	16-shot	48-shot
	In Context Learning	69.60			60.87			56.09		
TYPMarker	**Base**	59.77$_{\pm 2.90}$	66.04$_{\pm 1.34}$	74.91$_{\pm 0.38}$	66.29$_{\pm 2.19}$	79.77$_{\pm 1.61}$	80.85$_{\pm 1.30}$	54.42$_{\pm 4.56}$	65.33$_{\pm 4.43}$	83.10$_{\pm 0.47}$
	WSS	62.46$_{\pm 2.99}$	70.06$_{\pm 2.25}$	74.54$_{\pm 1.52}$	67.99$_{\pm 2.65}$	80.32$_{\pm 2.09}$	81.48$_{\pm 1.64}$	57.17$_{\pm 4.62}$	67.27$_{\pm 3.05}$	85.02$_{\pm 0.57}$
	WES	64.52$_{\pm 4.20}$	73.74$_{\pm 2.05}$	80.91$_{\pm 3.97}$	71.33$_{\pm 1.73}$	81.76$_{\pm 1.03}$	83.78$_{\pm 1.53}$	56.48$_{\pm 5.01}$	71.61$_{\pm 2.20}$	87.76$_{\pm 0.64}$
	LAMBADA	51.69$_{\pm 2.37}$	60.99$_{\pm 2.21}$	68.88$_{\pm 5.50}$	69.21$_{\pm 4.70}$	75.01$_{\pm 1.53}$	77.02$_{\pm 2.68}$	59.08$_{\pm 3.59}$	61.46$_{\pm 3.84}$	73.06$_{\pm 1.33}$
	ConsistRE	71.81$_{\pm 1.74}$	76.12$_{\pm 1.51}$	83.07$_{\pm 0.56}$	72.64$_{\pm 3.41}$	81.69$_{\pm 2.47}$	86.98$_{\pm 1.42}$	62.57$_{\pm 3.36}$	74.92$_{\pm 1.48}$	85.89$_{\pm 0.97}$
	UnleashLLMRE	69.81$_{\pm 2.55}$	75.00$_{\pm 1.41}$	81.49$_{\pm 0.96}$	72.41$_{\pm 1.47}$	81.56$_{\pm 1.63}$	86.16$_{\pm 1.86}$	62.80$_{\pm 3.87}$	73.01$_{\pm 2.36}$	84.41$_{\pm 0.11}$
	ODDA(Ours)	**75.78**$_{\pm 1.39}$	**80.48**$_{\pm 0.71}$	**84.91**$_{\pm 0.24}$	**76.78**$_{\pm 0.92}$	**84.86**$_{\pm 0.67}$	**88.89**$_{\pm 0.83}$	**67.57**$_{\pm 1.44}$	**78.93**$_{\pm 0.73}$	**88.99**$_{\pm 0.49}$
KnowPrompt	**Base**	62.08$_{\pm 3.91}$	67.39$_{\pm 1.97}$	76.57$_{\pm 0.96}$	52.02$_{\pm 5.99}$	74.47$_{\pm 0.19}$	80.98$_{\pm 2.46}$	51.85$_{\pm 5.98}$	67.99$_{\pm 1.37}$	78.64$_{\pm 1.19}$
	WSS	61.04$_{\pm 2.19}$	72.20$_{\pm 2.57}$	74.54$_{\pm 0.38}$	63.23$_{\pm 1.66}$	74.66$_{\pm 1.97}$	83.48$_{\pm 1.01}$	55.27$_{\pm 2.28}$	65.69$_{\pm 4.84}$	88.18$_{\pm 3.82}$
	WES	72.27$_{\pm 0.39}$	76.58$_{\pm 1.14}$	80.91$_{\pm 0.60}$	75.17$_{\pm 3.08}$	85.11$_{\pm 1.04}$	88.26$_{\pm 1.91}$	47.29$_{\pm 11.2}$	70.24$_{\pm 7.38}$	86.92$_{\pm 1.23}$
	LAMBADA†	58.45$_{\pm 1.85}$	69.75$_{\pm 0.54}$	68.88$_{\pm 1.61}$	52.71$_{\pm 2.87}$	68.12$_{\pm 1.77}$	81.22$_{\pm 0.83}$	48.41$_{\pm 5.14}$	65.00$_{\pm 3.68}$	79.01$_{\pm 4.02}$
	ConsistRE†	75.81$_{\pm 1.52}$	80.86$_{\pm 0.82}$	85.35$_{\pm 1.26}$	80.12$_{\pm 0.98}$	84.87$_{\pm 0.62}$	88.84$_{\pm 0.84}$	66.20$_{\pm 13.4}$	73.38$_{\pm 3.85}$	82.52$_{\pm 3.12}$
	UnleashLLMRE	74.68$_{\pm 1.26}$	78.32$_{\pm 1.32}$	83.40$_{\pm 0.40}$	80.49$_{\pm 4.10}$	84.47$_{\pm 1.87}$	88.86$_{\pm 1.31}$	67.76$_{\pm 4.14}$	74.54$_{\pm 0.60}$	84.31$_{\pm 2.36}$
	ODDA(Ours)	**80.96**$_{\pm 1.02}$	**84.82**$_{\pm 0.38}$	**87.02**$_{\pm 0.21}$	**84.12**$_{\pm 0.87}$	**87.94**$_{\pm 0.32}$	**90.08**$_{\pm 0.50}$	**71.48**$_{\pm 1.92}$	**77.86**$_{\pm 0.71}$	**88.64**$_{\pm 0.54}$

Table 2: Micro-F1 (%) across 8/16/48-shot settings. The best results are in **bold**, while the second-best ones are underlined. The green numbers indicate the minimum standard deviation from runs with different seeds. **Base** uses only the sampled seed data. **In-Context Learning** represents using sampled data as demonstrations for RE with LLMs. † indicates cases where the RE model fails to converge during training, highlighting the poor robustness of these RA methods. Performances from non-convergent models are excluded from the evaluation.

Method	TACRED			SemEval		
	Micro-F1	Self-BLEU	APS	Micro-F1	Self-BLEU	APS
ODDA(Ours)	75.78	0.4322	0.1189	67.57	0.2406	0.1042
Overall Diversity Data Selection (ODS)						
w/o ODS	73.29	0.4677	0.1315	65.76	0.2987	0.1177
w/o $\mathcal{R}^{inter}$	74.62	0.4463	0.1252	67.14	0.2420	0.1125
w/o $\mathcal{R}^{intra}$	74.15	0.4552	0.1248	66.87	0.2638	0.1098
Attribute-Constrained Data Generation (ACG)						
w/o ACG	-	0.5403	0.2266	-	0.4081	0.2202
w/o group	-	0.5316	0.1241	-	0.3125	0.1118
Selective Demonstration Filtering (SDF)						
w/o SDF	-	0.5695	0.1312	-	0.4520	0.1226

Table 3: Evaluating the influence of different parts in ODDA. 'w/o ODS' means randomly selecting the generated data to form the augmented dataset. 'w/o ACG' denotes generating based solely on schema constraints. 'w/o group' refers to randomly selecting demonstrations for each instance. 'w/o SDF' indicates not selecting ambiguous data as demonstrations.

continuous tuning with synergistic optimization. More details are in Appendix C.

5.2 Main Results

In Table 2, we present the Micro-F1 scores and their standard deviation over 5 runs in three datasets. The use of DA results in enhanced performances by utilizing augmented data, which also outperforms the ICL-based RE. This highlights the effectiveness of DA for RE and points to the limitations of relying solely on LLMs for complex RE tasks (achieve $\sim 60\%$ F1 but require billions of parameters).

In low-resource scenarios, the proposed method consistently outperforms others (more than 1%↑) across various settings. It achieves particularly notable improvements (about 5% ↑) when fewer sampled data (*e.g.* 8-shot) are used. Additionally, the proposed methods contribute to more stable performances, as evidenced by the minimal standard deviations (most less than 1%).

Furthermore, the proposed method shows significant performance improvements across different types of RE base models, including fine-tuning-based and prompt-tuning-based models. Compared to other DA methods, our method delivers greater enhancements, demonstrating strong applicability.

5.3 Ablation Study

Given the similar results between TACRED and ReTACRED, we conduct experiments on the more discriminative 8-shot setting using TACRED and SemEval datasets. The results of removing different components from the proposed method are presented in Table 3. The results indicate that ODS effectively identifies data with greater overall diversity from the large pool of generated data. A significant performance drop would occur if data are selected purely at random. Notably, the inter-class rewards demonstrate a slightly greater impact on diversity and RE performance compared to intra-class rewards, highlighting the importance of selecting diverse instances from a global perspective. ACG not only enhances structural diversity but also significantly improves semantic diversity, underscoring the critical role of attribute constraints in the generation process. Furthermore, SDF contributes significantly to enhancing the diversity of the generated data. This finding emphasizes that

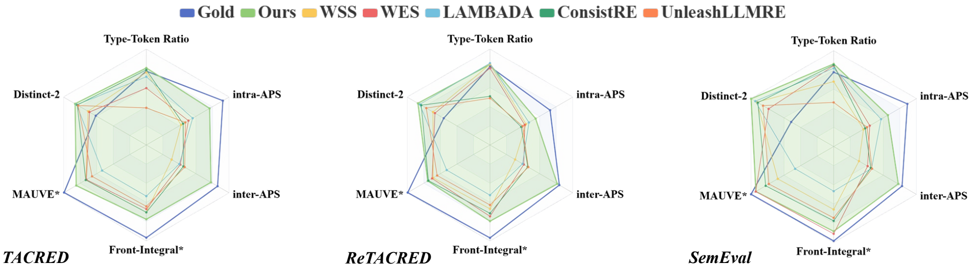


Figure 3: Comparison of methods regarding data diversity and quality on three datasets. The metrics include Type-Token Ratio (↑), Distinct-*N* (↑), and MAUVE* (↑), as well as inter-APS (↓), intra-APS (↓), and Front-Integral* (↓). For unified visualization, the Front-Integral* is plotted as 1 - *value*, while inter-APS and intra-APS are represented as 1 / *value*. The **blue line** represents the `Gold` data and **green area** represents the proposed method.

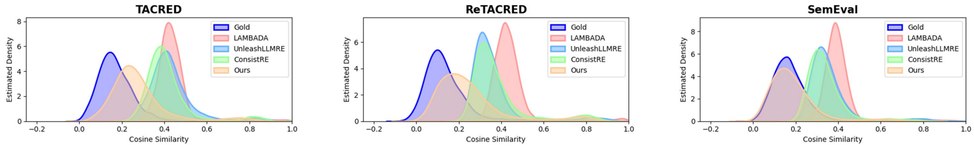


Figure 4: The distribution of cosine similarity of sentence pairs sampled from the same relation type.

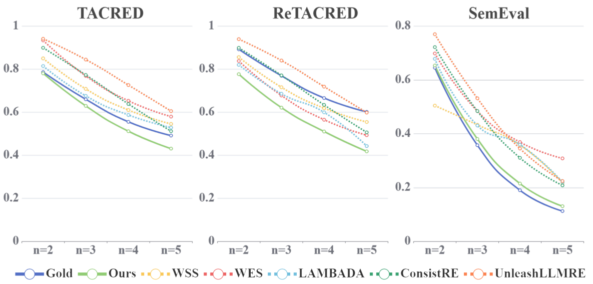


Figure 5: Self-BLEU (↓) for *n*-grams across different datasets. We calculate based on 25,000 generated samples for TACRED and ReTACRED, and 6,500 for SemEval by each method to evaluate the overall diversity.

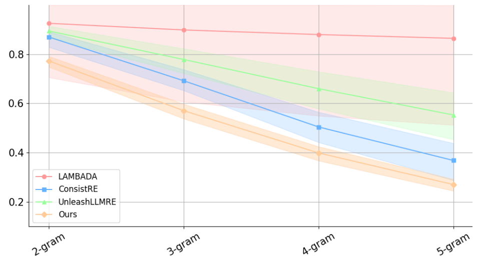


Figure 6: Self-BLEU for the augmented data on TACRED with 5- to 100-shot. The lines represent the mean Self-BLEU of different methods, while the shaded areas indicate the range of *min* to *max* under different shots.

understanding the model's behavior provides insights into the structural and semantic types of data it requires. Consequently, data generated based on ambiguous instances allows the model to focus on the most relevant and diverse instances, thereby improving overall effectiveness in the RE task.

5.4 Discussion of Data Diversity

Impact on Overall Sample Diversity. We comprehensively evaluate the diversity of augmented data from different DA methods through multi-dimensional analysis. As illustrated in Figure 3, the proposed method outperforms others in both diversity and quality, demonstrating its effectiveness in generating diverse samples for RE tasks. Its reader plot area closely aligns with the Gold data and even surpasses it in Distinct-2.

For lexical diversity, an interesting observation is that LLM-based DA methods produce more unique bigrams but achieve lower TTR. Following the previous work (Hu et al., 2023), we compute TTR based on the dependency path between head and tail entities. It reveals that LLM-based methods tend to introduce variability in sentence parts unrelated to target triplets, contributing minimally to performance improvement. Self-BLEU scores, as depicted in Figure 5, further highlight lexical diversity across methods by assessing n-gram overlaps. ODDA achieves lower Self-BLEU, indicating higher inter-sample diversity. While ConsistRE despite explicit diversity-enhancing designs, achieves higher scores, reflecting limited overall diversity.

For semantic diversity, measured by APS, existing methods show higher intra/inter-class APS, indicating greater sample homogeneity. They focus on individual textual variations without addressing inter-sample diversity, resulting in homogeneous yet less diverse datasets. Figure 4 validates this point by visualizing the cosine similarity distribution of same-relation text pairs. Traditional methods sacrifice diversity through token substitutions. LLM-based methods are constrained by the LLM's intrinsic biases, resulting in sample homogeneity. ODDA achieves distributions closest to Gold data, underscoring its ability to generate diverse samples.

Finally, the proposed method ensures both sample diversity and high-quality generation, as evidenced by its MAUVE* and Front-Integral* scores in Figure 3, which align closely with Gold data. This superior performance highlights its capability to produce diverse, human-like samples.

Impact of Different Shot on Sample Diversity. In Figure 6 and Appendix D.1, we investigate the data diversity under different settings, a factor overlooked in previous work. Notably, we observe that WSS, WES, and LAMBADA share a similar trend. Only the results of LAMBADA are presented for clarity. It reveals that the sample diversity of existing methods changes dramatically when the number of available samples varies (*e.g.* 5/100-shot), as reflected in the size of the shaded areas. Interestingly, methods designed to boost diversity may sometimes yield less diversity than simple word-substitution strategies, as reflected by the overlapping shaded region. This occurs because such methods focus on per-sample variations and thus rely on sampling more examples to achieve broader diversity. By contrast, ODDA adapts well to varying settings, producing consistently diverse samples.

Impact on Sample Distribution. We further verify the diversity of generated data and whether it meets the requirements of the target model through feature visualization, as shown in Figure 7 and Appendix D.2. ODDA generates diverse data from an overall view, ensuring a relatively uniform distribution and better alignment with target data. It reveals that the data generated by other methods exhibits clear boundaries with the target distribution and such domain shift may decrease the model performance (Divekar and Durrett, 2024). Additionally, traditional methods tend to generate data similar to the seed data. While LLM-based methods produce similar content when given analogous seed data, as highlighted by the red box. The data diversity falls into a local optimum, harming RE performance.

5.5 Discussion of Data Quantity

Impact of Generated Data Size. Previous research (Ye et al., 2022; Xu et al., 2023b) indicates that generating more data may not improve RE performance. As shown in Figure 8 and Appendix D.3, existing methods experience fluctuations in performance with an increasing quantity of generated data, accompanied by a substantial degree of uncertainty (even variance > 5). This limitation arises because they provide data without considering the overall distribution, which hinders data diversity and fails to provide effective assistance to the model. In contrast, the proposed method tailors the generated data to the target model and selects data from a global perspective. This allows for the effective utilization of data, leading to saturation requires more data. Finally, ODDA achieves great performance improvements and stability.

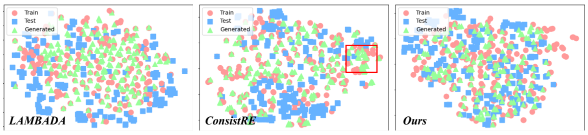


Figure 7: Visualization for the original, augmented, and test datasets after using t-SNE dimension reduction.

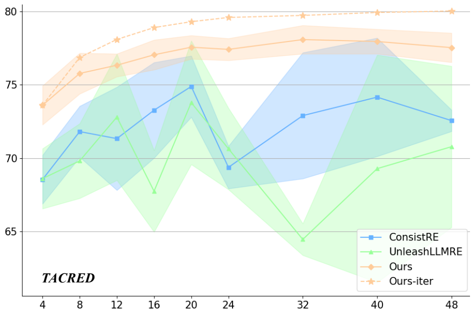


Figure 8: Micro-F1 scores at different scaling factors under the seed data from TACRED. The shaded areas represent the variance across multiple runs. 'Ours-iter' denotes generating data iteratively, where each iteration restarts the OODA loop to generate augmented data and achieve the current scaling factor.

Generate Data Once or Iteratively. As shown in Figure 8 and Appendix D.3, rather than generating the target augment ratio at once (Ours), we iteratively restart the OODA loop to provide an additional 4 or 8× more data (Ours-iter). This strategy accelerates the performance improvement of the model, as the proposed method selects the most valuable data based on the observed behavior of the target RE model and then generates diverse data. Furthermore, through multiple cycles of iterative optimization, the RE model achieves enhanced performance (2%↑) under the same k-shot setting.

6 Conclusion

In this paper, we propose a novel DA method for LRE, called ODDA, that integrates three key components: SDF, ACG, and ODS. Through extensive experiments, ODDA consistently outperforms existing methods in performance, stability, and data diversity. This work offers a novel deconstruction of augmented data generation and selection, providing new insights into generating diverse and effective data for LRE and other NLP tasks.

Limitation.

Despite the proposed ODDA achieving significant results, there remain several limitations and areas for improvements that provide avenues for future investigation while not affecting the overall integrity and innovation of our contributions.

Task Settings. This study is centered on DA for sentence-level relation extraction in low-resource scenarios. It limits its direct applicability to more complex tasks such as document-level relation extraction (DocRE), Event Extraction (EE), Named Entity Recognition (NER), and other structured prediction tasks. Although it remains an open challenge to adopt a unified paradigm to resolve all complex settings, the proposed ODDA has promising adaptive capabilities to handle these tasks. In future work, we intend to explore adjustments to the Data Map computation, attributes design, and reward function, with the objective of constructing a more versatile multi-task DA system.

Privacy. Secondly, the proposed ODDA does not involve the processing of personally identifiable information and consequently does not support data desensitization. Scenarios requiring the handling of privacy data require alternative approaches. In such cases, substituting API-based LLM calls with an offline deployment could effectively serve as a safeguard to prevent data leakage in ODDA.

Ethics Statement. Additionally, data generated by LLMs may contain biased sentences that could raise ethical concerns. Such outputs do not reflect the views of the authors. Additionally, if the proposed method encounters adversarial or offensive input, ODDA itself does not filter the content, as this falls outside the scope of this study. However, it is worth noting that current LLMs are equipped with safety measures to detect high-risk inputs and restrict refuse to generate corresponding outputs.

7 Acknowledgement

This work is supported by the National Natural Science Foundation of China (62176185, U23B2057).

References

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! In AAAI, pages 7383–7390. AAAI Press.

Hengyi Cai, Hongshen Chen, Yonghao Song, Cheng Zhang, Xiaofang Zhao, and Dawei Yin. 2020. Data manipulation: Towards effective instance learning for neural dialogue generation via learning to augment and reweight. In ACL, pages 6334–6343. Association for Computational Linguistics.

Huajun Chen. 2024. Large knowledge model: Perspectives and challenges. Data Intelligence, 6(3):587–620.

Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In WWW, pages 2778–2788. ACM.

Shumin Deng, Yubo Ma, Ningyu Zhang, Yixin Cao, and Bryan Hooi. 2024. Information extraction in low-resource scenarios: Survey and perspective. In ICKG. 15th IEEE International Conference on Knowledge Graphs.

Abhishek Divekar and Greg Durrett. 2024. Synthesizrr: Generating diverse datasets with retrieval augmentation. In EMNLP, pages 19200–19227. Association for Computational Linguistics.

Alexander R. Fabbri, Simeng Han, Haoyuan Li, Haoran Li, Marjan Ghazvininejad, Shafiq R. Joty, Dragomir R. Radev, and Yashar Mehdad. 2021. Improving zero and few-shot abstractive summarization with intermediate fine-tuning and data augmentation. In NAACL-HLT, pages 704–717. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Yijie Zhong, Yuxi Bi, Ming Xue, and Haofen Wang. 2025. Synergizing RAG and reasoning: A systematic review. CoRR, abs/2504.15909.

Sireesh Gururaja, Ritam Dutt, Tinglong Liao, and Carolyn P. Rosé. 2023. Linguistic representations for fewer-shot relation extraction across domains. In ACL (1), pages 7502–7514. Association for Computational Linguistics.

Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. 2023. Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors. CoRR, abs/2305.14450.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and

Stan Szpakowicz. 2010. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In SemEval@ACL, pages 33–38. The Association for Computer Linguistics.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2019. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. CoRR, abs/1911.10422.

Xuming Hu, Aiwei Liu, Zeqi Tan, Xin Zhang, Chenwei Zhang, Irwin King, and Philip S. Yu. 2023. GDA: generative data augmentation techniques for relation extraction tasks. In ACL (Findings), pages 10221–10234. Association for Computational Linguistics.

Xuming Hu, Chenwei Zhang, Fukun Ma, Chenyao Liu, Lijie Wen, and Philip S. Yu. 2021. Semi-supervised relation extraction via incremental meta self-training. In EMNLP (Findings), pages 487–496. Association for Computational Linguistics.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling BERT for natural language understanding. In EMNLP (Findings), volume EMNLP 2020 of Findings of ACL, pages 4163–4174. Association for Computational Linguistics.

Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. 2023a. Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. CoRR, abs/2304.11633.

Guozheng Li, Peng Wang, and Wenjun Ke. 2023b. Revisiting large language models as zero-shot relation extractors. In EMNLP (Findings), pages 6877–6892. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In HLT-NAACL, pages 110–119. The Association for Computational Linguistics.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022a. WANLI: worker and AI collaboration for natural language inference dataset creation. In EMNLP (Findings), pages 6826–6847. Association for Computational Linguistics.

Fangchao Liu, Hongyu Lin, Xianpei Han, Boxi Cao, and Le Sun. 2022b. Pre-training to match for unified low-shot relation extraction. In ACL (1), pages 5785–5795. Association for Computational Linguistics.

Lang Liu, Krishna Pillutla, Sean Welleck, Sewoong Oh, Yejin Choi, and Zaid Harchaoui. 2021. Divergence Frontiers for Generative Models: Sample Complexity, Quantization Effects, and Frontier Integrals. In NeurIPS.

David Lowell, Brian E. Howard, Zachary C. Lipton, and Byron C. Wallace. 2021. Unsupervised data augmentation with naive augmentation and without unlabeled data. In EMNLP (1), pages 4992–5001. Association for Computational Linguistics.

Mingyu Derek Ma, Xiaoxuan Wang, Po-Nien Kung, P. Jeffrey Brantingham, Nanyun Peng, and Wei Wang. 2024. STAR: boosting low-resource information extraction by structure-to-text data generation with large language models. In AAAI, pages 18751–18759. AAAI Press.

Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic data augmentation increases robustness to inference heuristics. In ACL, pages 2339–2352. Association for Computational Linguistics.

Swaroop Mishra, Anjana Arunkumar, Bhavdeep Singh Sachdeva, Chris Bryan, and Chitta Baral. 2020. DQI: measuring data quality in NLP. CoRR, abs/2005.00816.

Francesco Molfese, Simone Conia, Riccardo Orlando, and Roberto Navigli. 2024. ZEBRA: zero-shot example-based retrieval augmentation for commonsense question answering. In EMNLP, pages 22429–22444. Association for Computational Linguistics.

Jonas Mueller and Aditya Thyagarajan. 2016. Siamese recurrent architectures for learning sentence similarity. In AAAI, pages 2786–2792. AAAI Press.

Rosel Oida-Onesa and Melvin A. Ballera. 2024. Fine tuning language models: A tale of two low-resource languages. Data Intelligence, 6(4):946–967.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In ACL, pages 311–318. ACL.

Krishna Pillutla, Lang Liu, John Thickstun, Sean Welleck, Swabha Swayamdipta, Rowan Zellers, Sewoong Oh, Yejin Choi, and Zaid Harchaoui. 2023. MAUVE Scores for Generative Models: Theory and Practice. JMLR.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. In NeurIPS.

Injy Sarhan and Marco Spruit. 2020. Can we survive without labelled data in nlp? transfer learning for open information extraction. Applied Sciences, 10(17).

George Stoica, Emmanouil Antonios Platanios, and Barnabás Póczos. 2021. Re-tacred: Addressing shortcomings of the TACRED dataset. In AAAI, pages 13843–13850. AAAI Press.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In EMNLP (1), pages 9275–9293. Association for Computational Linguistics.

Fiona J. Tweedie and R. Harald Baayen. 1998. How variable may a constant be? measures of lexical richness in perspective. Comput. Humanit., 32(5):323–352.

Amir Pouran Ben Veyseh, Franck Dernoncourt, Bonan Min, and Thien Huu Nguyen. 2023. Generating labeled data for relation extraction: A meta learning approach with joint GPT-2 training. In ACL (Findings), pages 11466–11478. Association for Computational Linguistics.

Jason W. Wei and Kai Zou. 2019. EDA: easy data augmentation techniques for boosting performance on text classification tasks. In EMNLP/IJCNLP (1), pages 6381–6387. Association for Computational Linguistics.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark O. Riedl, and Yejin Choi. 2022. Reframing human-ai collaboration for generating free-text explanations. In NAACL-HLT, pages 632–658. Association for Computational Linguistics.

Benfeng Xu, Quan Wang, Yajuan Lyu, Dai Dai, Yongdong Zhang, and Zhendong Mao. 2023a. S2ynre: Two-stage self-training with synthetic data for low-resource relation extraction. In ACL (1), pages 8186–8207. Association for Computational Linguistics.

Xin Xu, Xiang Chen, Ningyu Zhang, Xin Xie, Xi Chen, and Huajun Chen. 2022. Towards realistic low-resource relation extraction: A benchmark with empirical baseline study. In EMNLP (Findings), pages 413–427. Association for Computational Linguistics.

Xin Xu, Yuqi Zhu, Xiaohan Wang, and Ningyu Zhang. 2023b. How to unleash the power of large language models for few-shot relation extraction? In SustaiNLP, pages 190–200. Association for Computational Linguistics.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. Zerogen: Efficient zero-shot learning via dataset generation. In EMNLP, pages 11653–11669. Association for Computational Linguistics.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J. Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large language model as attributed training data generator: A tale of diversity and bias. In NeurIPS.

Xilin Zhang, Zhixin Mao, Ziwen Chen, and Shen Gao. 2024. Effective tool augmented multi-agent framework for data analysis. Data Intelligence, 6(4):923–945.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In EMNLP, pages 35–45. Association for Computational Linguistics.

Yifan Zheng, Wenjun Ke, Qi Liu, Yuting Yang, Ruizhuo Zhao, Dacheng Feng, Jianwei Zhang, and Zhi Fang. 2024. Making llms as fine-grained relation extraction data augmentor. In IJCAI, pages 6660–6668. ijcai.org.

Yijie Zhong, Feifan Wu, Mengying Guo, Xiaolian Zhang, Meng Wang, and Haofen Wang. 2025. Meta-pke: Memory-enhanced task-adaptive personal knowledge extraction in daily life. Inf. Process. Manag., 62(4):104097.

Wenxuan Zhou and Muhao Chen. 2022. An improved baseline for sentence-level relation extraction. In AACL/IJCNLP (2), pages 161–168. Association for Computational Linguistics.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In SIGIR, pages 1097–1100. ACM.

Tianyuan Zou, Yang Liu, Peng Li, Jianqing Zhang, Jingjing Liu, and Ya-Qin Zhang. 2024. Fusegen: PLM fusion for data-generation based zero-shot learning. In EMNLP, pages 2172–2190. Association for Computational Linguistics.

A Details of Datasets

TACRED is a large-scale crowd-sourced relation extraction dataset that serves as a challenging benchmark due to its diverse relations and complex language. It contains 42 relation types, including 'no_relation', meaning no relation is found. It is collected from previous TAC KBP shard tasks. TACRED is available on the official website [3].

ReTACRED is a re-annotated version of TACRED with 40 relation types. It contains 39 common relation types and 1 special 'NA' relation type, meaning none of the above. ReTACRED can be obtained from the website [4].

SemEval is a traditional dataset widely employed in relation extraction. It is a human-annotated dataset from SemEval-2010 Task 8 (Hendrickx et al., 2010) and is devoid of noise. It contains 19 relation types: Cause-Effect, Component-Whole, and Others. SemEval is an open-source resource that is publicly accessible [5].

[3] https://nlp.stanford.edu/projects/tacred

[4] https://github.com/gstoica27/Re-TACRED

[5] https://huggingface.co/datasets/SemEvalWorkshop

B Details of Metrics

B.1 Type-Token Ratio

Following the previous work (Hu et al., 2023), we introduce the Type-Token Ratio (TTR) to measure the diversity of augmented sentences. TTR measures the ratio of the number of different words to the total number of words in the dependency path between the given head entity and tail entity. Higher TTR (%) indicates more diversity in sentences. It can be calculated as follows:

$$TTR = \frac{|\mathbb{V}(\bigcup_{(s,r,h,t)\in D} P_{h\to t})|}{\sum_{(s,r,h,t)\in D}|P_{h\to t}|}, \qquad (4)$$

where $P_{h\to t}$ represents the words in the dependency path between the head entity and the tail entity, and $\mathbb{V}(\cdot)$ represents the unique words among all the words.

B.2 Distinct-N

Following the previous work (Zheng et al., 2024), we introduce the Distinct (Li et al., 2016) to assess the diversity of synthetic sentences. Distinct qualifies the number of distinct unigrams and bigrams divided by the total number of generated words. The calculation formula is as follows:

$$Distinct(N) = \frac{\text{Unique N-grams}}{\text{Total N-grams}} \times 100\%. \qquad (5)$$

In this paper, we set $N = 2$, representing the proportion of unique bigrams. Higher Distinct-N indicates more diversity in sentences.

B.3 Self-BLEU

Since *TTR* and *Distinct* do not take into account the similarity between sentences, we introduce Self-BLEU (Zhu et al., 2018) to measure lexical diversity of the augmented dataset based on *n*-gram overlap between pairs of sentences. Self-BLEU helps ensure that the augmented instances introduce meaningful variability into the dataset, avoiding redundancy and enhancing the model's ability to generalize across diverse linguistic expressions of the same relation. It can be calculated as follows:

$$\text{Self-BLEU}(D) = \frac{1}{|D|}\sum_{s\in D}\text{BLEU}(s, D\setminus\{s\}), \qquad (6)$$

where BLEU indicates the *naïve* BLEU score (Papineni et al., 2002). High Self-BLEU (close to 1) indicates low diversity, suggesting that the generated instances are highly similar to each other. Low Self-BLEU (closer to 0) indicates high diversity, reflecting a broader range of linguistic patterns and lexical choices in the generated data. When *n* ranges from 2 to 5, it captures both lexical and syntactic diversity. For computational efficiency, we employ *fast-bleu* [6][7] in this paper.

B.4 Average Pairwise Sample Similarity

While the aforementioned metrics focus solely on structural diversity, we introduce Average Pairwise Sample Similarity (APS) (Yu et al., 2023) to assess semantic diversity. APS measures the average similarity between pairs of sentences based on their semantic embeddings, providing a quantitative evaluation of how diverse the generated sentences are in terms of meaning. Furthermore, to gain deeper insights, we compute both intra-class APS and inter-class APS, respectively capturing diversity within the same relation type and across different relation types. Importantly, a lower APS value indicates better diversity, as it signifies reduced semantic overlap among sentences.

B.5 MAUVE and Front-Integral

We aim to provide a comprehensive evaluation that captures both the semantic alignment with human-written text and the balance between diversity and quality in the generated data. Therefore, we introduce MAUVE (Pillutla et al., 2021) and Front-Integral (Liu et al., 2021) to evaluate the quality of generated sentences. They address the limitations of traditional metrics like BLEU or ROUGE and offer a more reliable measure of distributional similarity in a global view.

MAUVE quantifies the similarity between the distributions of generated and golden data by leveraging embeddings in a low-dimensional space. It computes the divergence between these distributions using a combination of Kernel Density Estimation (KDE) and Optimal Transport (OT), providing a single scalar score that reflects the alignment between the two distributions. A higher MAUVE score indicates better agreement, suggesting that the generated sentences are more human-like.

Front-Integral constructs a Pareto frontier between these two dimensions and computes the area under the curve, representing the overall performance of the generated sentences. A lower Front-integral indicates higher-quality generated data.

[6] https://github.com/Danial-Alh/fast-bleu

[7] github.com/yizhangliu/fast-bleu_windows_vs2019

In this paper, we employ the improved version computed with Krichevsky-Trofimov smoothing (Pillutla et al., 2023), *i.e.* MAUVE* and Front-Itegral*, and use the official code [8]. Note that these metrics are suited for relative comparisons while the absolute scores are less meaningful.

B.6 Sampling Size for Metric Computation

It is worth noting that this work's analysis focuses on relative metric differences rather than absolute values. We calculate the metrics for TTR and Distinct-N on 1,024 samples. Following official recommendations, we calculate MAUVE$star$ and Front-Integral$star$ed on 4,096 samples. In terms of Self-BLEU, we utilize 25,000 samples on TACRED/ReTACRED and 6,500 samples on SemEval (constrained by dataset size) for comprehensive DA method assessment. We utilize 5,000 samples for varying low-resource settings, considering equitable evaluation across extreme data scarcity levels (**i.e.** 5-shot). Samples are randomly selected from all the augmented data from different DA methods.

C Implementation Details

C.1 Preparation of k-shot Seed Data

In this paper, we follow the previous research (Xu et al., 2022) for k-shot data sampling. Specifically, we randomly select k instances to constitute the k-shot dataset across every relation type within each RE dataset. If a particular relation type contains fewer than k instances, we utilize all available data for that relation type. We leverage the open-source code [9] provided in the existing work. To ensure robustness and minimize randomness bias, we conduct this sampling process five times, each initiated with a distinct random seed, thus yielding five unique k-shot datasets.

C.2 Implementation of the base RE model

All training processes for the base RE models are conducted on a single 24GB NVIDIA 3090 GPU. For the LLM-based RE with In-Context Learning, we utilize DeepSeek-v3 [10] as the base LLM.

TYPMarker (Zhou and Chen, 2022) is an improved RE baseline and adopts the Typed Entity Marker (punct) technique. This method enhances entity representation by marking entity spans and their types without introducing new special tokens. Specifically, the head entity is enclosed with ‘@’ and its type is prepended with ‘$*$’, while the tail entity is enclosed with ‘#’ and its type is prepended with ‘$\wedge$’. The embeddings for these markers are randomly initialized and fine-tuned during training. These embeddings are then fed into the classifier to output the predicted probability distribution for the pre-defined relations. We use the open-source code [11] and make the super-parameters unchanged. We apply the RoBERTa-large as the backbone.

KnowPrompt (Chen et al., 2022) is a knowledge-aware prompt-tuning framework for RE. It introduces virtual answer words and virtual type words to encode semantic knowledge from relation labels and entity types, respectively. These virtual words are synergistically optimized with context-aware prompt calibration and implicit structured constraints, ensuring they adapt to the surrounding context and maintain relational semantics. This approach allows the model to effectively leverage task-specific knowledge without extensive fine-tuning, making it particularly effective in both standard and low-resource settings. We utilize the official provided code [12] and super-parameters. Specifically, we apply RoBERTa-large as the backbone.

In-Context Learning (ICL) leverages the powerful generative capabilities of LLMs to perform relation extraction in a few-shot or zero-shot manner. Demonstrations serve as ICL prompts, enabling the model to better understand and generalize the semantic patterns of the specific relation. In this paper, we use the open-source code in previous work (Xu et al., 2023b) for experimental validation. The detailed prompt is shown as follows:

> Prompt: ICL for RE
>
> Given a context, a pair of head and tail entities in the context, decide the relationship between the head and tail entities from candidate relations: <relation type 1>, <relation type 2>, ...
> Context: <sentence>. The relation between <head entity> and <tail entity> in the context is <relation type>.
> Context: ...
> Context: <query sentence>. The relation between <query head entity> and <query tail entity> in the context is

The parameter *temperature* is set to 0 for precision in ICL. For each relation type, we provide one example that demonstrates the relation in the context. Finally, the LLM's response is post-processed

[8] https://github.com/krishnap25/mauve

[9] https://github.com/zjunlp/DeepKE

[10] https://platform.deepseek.com

[11] https://github.com/wzhouad/RE_improved_baseline

[12] https://github.com/zjunlp/KnowPrompt

through regular expression matching to extract and validate the predicted relation types.

C.3 Implementation of the Compared Method

We generate 48 augmented data points for each seed data point across all DA methods to conduct extensive experiments, from which 8 are carefully selected for the main experiment.

Non-LLM based DA methods. In this paper, we leverage the open-source code [13] and primarily utilize the *nlpaug* library [14] to implement the WES, WSS, and LAMBADA. Specifically, we employ the *GoogleNews-vectors* for WES, while the WSS is implemented with *RoBERTa-large*. For LAMBADA, we conduct separate training tailored to each dataset, ensuring optimal performance and adaptability to the unique dataset.

LLM based DA methods. In our experiments, we apply the *DeepSeek-v3* for the LLM-based DA methods. The *temperature* parameter is set to 2. For ConsistRE, we utilized the prompt templates described below.

Prompt: ConsistRE

> Knowledge: The relation between <head entity> and <tail entity> is <relation type>
> Objective: Make sentences with given entities <head entity>, <tail entity> and keyword <keyword hint>
> Output: <sentence>
> Knowledge: ...
> Objective: ...
> Output: ...
> Knowledge: The relation between <target head entity> and <target tail entity> is <target relation type>
> Objective: Make sentences with given entities <target head entity>, <target tail entity> and keyword <target keyword hint>
> Output:

For UnleashLLMRE, we utilize the original open-source code [15], organizing the provided prompt templates as below.

Prompt: UnleashLLMRE

> One sample in relation extraction datasets consists of a relation, a context, and a pair of head and tail entities in the context.
> The head entity has a relation with the tail entity.
> Here are some samples for relation '<relation type>':
> Relation: <relation type>. Context: <sentence>. Head Entity: <head entity>. Tail Entity: <tail entity>.
> Relation: ... Context: ... Head Entity: ... Tail Entity: ...
> Generate <number> samples for the relation '<target relation type>', head entity '<target head entity>', and tail entity '<target tail entity>'.

C.4 Implementation of the Proposed Method

Selective Demonstration Filtering. We train the target RE model for 300 steps, recording the *logits* of the gold relation type for all the available data every 20 steps. This entire process is highly efficient and completed rapidly. Subsequently, we compute the variability and confidence scores for each sample based on the recorded *logits*. To identify ambiguous samples, we select data points where both variability and confidence fell within the range of $[0.3, 0.7]$ (*i.e.* $\tau = 0.3$ and $\epsilon = 0.7$).

Attribute Constrained Data Generation. To ensure the rationality and comprehensiveness of key attributes, we follow the existing research (Yu et al., 2023) and adopt a human-machine collaboration strategy to determine the generation conditions required for generating diverse samples. Specifically, we query the LLMs with the following content.

Prompt: Query for Generation Conditions

> When trying to generate multiple sentences that all express the same specific relation (i.e., same head entity–relation–tail entity) yet exhibit diversity, it helps to think about the text-level "conditions" you can vary. Key conditions often include:

Through manual validation and selection of the LLM's responses, the demonstration provision includes four key attributions: relation type, entity, semantic, and syntax. The generation conditions consist of six attributes: writing style, expression of relations, sentence order and length, voice and perspective, tense, and tones and sentiment. The complete prompt is shown in Table 4 for reference.

Overall Diversity Date Selection. In the implementation of Monte Carlo Tree Search (MCTS), at each step, one data point is selected from each relation category to initiate the search. The search process terminates when the search space is exhausted

[13] https://github.com/zjunlp/LREBench
[14] https://github.com/makcedward/nlpaug
[15] https://github.com/zjunlp/DeepKE/tree/main/example/llm/UnleashLLMRE

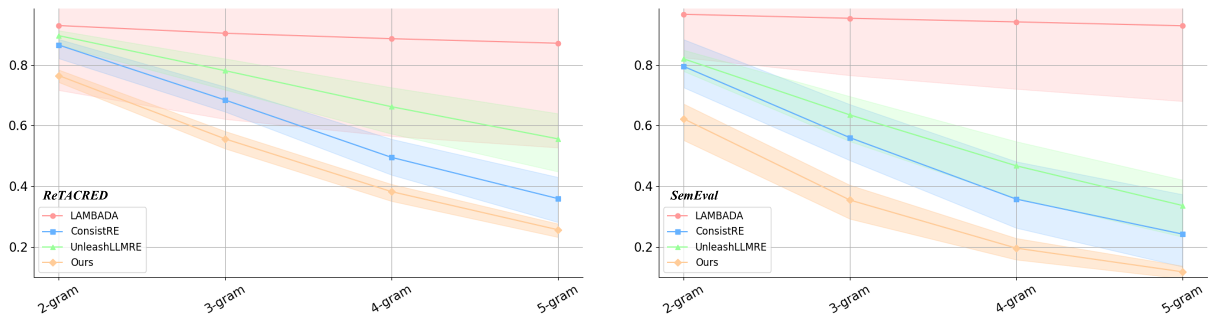


Figure 9: Self-BLEU scores for the augmented data on ReTACRED and SemEval with 5-shot to 100-shot. The lines represent the mean Self-BLEU of different methods, while the shaded areas indicate the range of *min* to *max* under different shots.

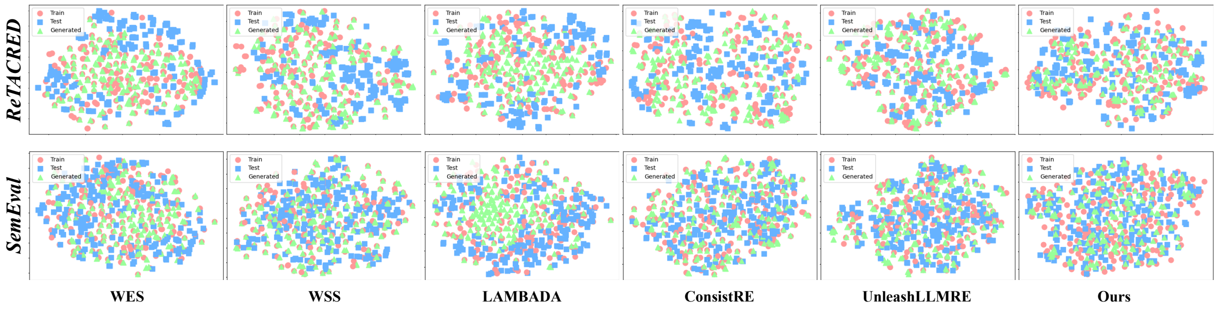


Figure 10: Visualization for the original, test, and augmented datasets from different DA methods after using t-SNE dimension reduction.

or when the number of simulations reaches 100. We employ the Upper Confidence Bound for Trees (UCT) search strategy, which can be expressed by the following formula:

$$UCT = \frac{\mathcal{R}(n')}{N(n')} + 2 \cdot \sqrt{\frac{\ln N(n)}{N(n')}}, \qquad (7)$$

where n and n' represent a parent node in the search tree and its child node respectively. $\mathcal{R}(\cdot)$ means the reward accumulated by a node and $N(\cdot)$ tracks the number of visits to a node.

D More Experimental Results

D.1 Impact of Different Shot on ReTACRED and SemEval

Figure 9 showcases the Self-BLEU for augmented data on ReTACRED and SemEval datasets across various n-grams. The results reveal that the proposed method consistently achieves the lowest Self-BLEU scores across all n-grams on both datasets, signifying its ability to produce highly diverse augmented data. In contrast, LAMBADA (as well as WSS, and WES, which have similar results) exhibits the highest Self-BLEU, indicating limited diversity and the generation of numerous similar samples. For ConsistRE and UnleashLLMRE, while they show better diversity compared to LAMBADA, their Self-BLEU remains notably higher than those of our approach, especially for higher n-grams (4-gram and 5-gram). Moreover, the shaded areas representing score ranges indicate that the proposed method maintains consistently low variation, underscoring the stability and reliability of its data generation process.

D.2 Impact of Sample Distribution on ReTACRED and SemEval

Figure 10 further visualizes the sample distribution of the original (orange), test (blue), and augmented (green) datasets after applying t-SNE dimensionality reduction for ReTACRED and SemEval. Across all DA methods, noticeable differences in the distribution patterns can be observed. For the WES, WSS, and LAMBADA, we observe that the generated data distributions closely overlap with the original data, maintaining a distinct boundary from the test data. This suggests limited diversity in their augmented datasets. For ConsistRE and UnleashLLMRE, the introduction of LLMs improves data diversity, resulting in augmented data distributions that align more closely with the test data. However, they still suffer from generating similar

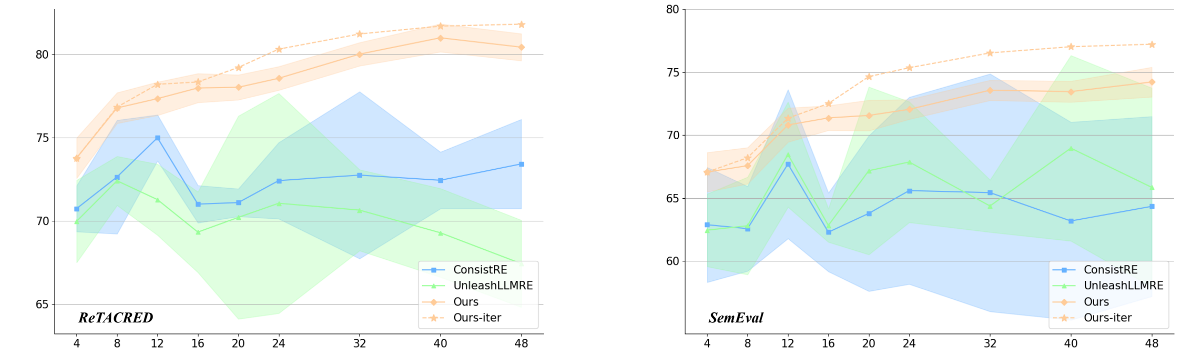


Figure 11: Micro F1 scores at different scaling factors under the seed data from ReTACRED and SemEval. The shaded areas represent the variance across multiple runs. 'Ours-iter' denotes generating data iteratively, where each iteration restarts the OODA loop to generate augmented data and achieve the current scaling factor.

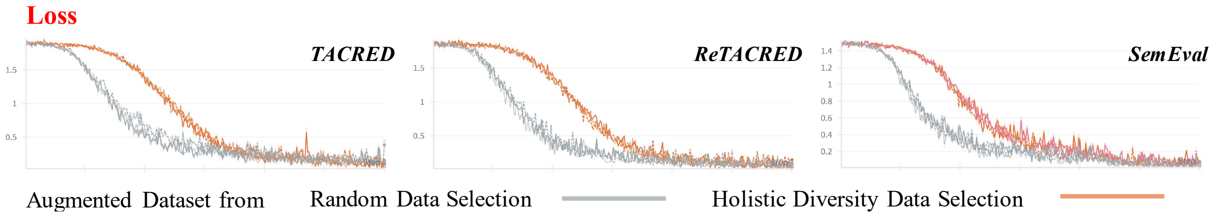


Figure 12: Training loss comparing different data selection strategies for augmented data. The proposed overall diversity data selection generates a diverse augmented dataset, which effectively facilitates model training. In contrast, random selection reduces the diversity of augmented data, causing the model to repeatedly encounter similar samples and accelerating overfitting.

samples across different inputs, as evidenced by the aggregation of green points in the visualizations. In contrast, the proposed method achieves both greater overall diversity and better alignment with the target distribution. The augmented data maintains a distinct separation from the original data while also exhibiting variance among themselves.

D.3 Impact of Generated Data Size on ReTACRED and SemEval

Figure 11 illustrates the RE performances at various scaling factors for ReTACRED and SemEval datasets. Similar to the results observed on the TACRED dataset, the proposed method demonstrates consistently better and more stable performance across various scaling factors. It effectively utilizes more generated data, whereas other methods often experience performance degradation as the data volume increases. Notably, on the SemEval dataset, certain existing methods occasionally show better performance. However, these instances are attributed to the randomness, as performance could fluctuate by more than 10% when different random seeds are used. Furthermore, with repeated restarts of the OODA loop in the proposed method, we observe significant performance improvements as the data volume increases. This highlights the ability of the iterative process to generate useful augmented data, ensuring diversity and alignment with the target RE model's requirement, ultimately achieving superior performance.

D.4 Discussion on Training Loss Trend

During the training process, we observed several notable phenomena reflected in Figure 12. For all three datasets, the overall diversity data selection strategy (orange) led to a more gradual and consistent reduction in training loss compared to random data selection (gray). This suggests that the diverse augmented dataset generated by our method helps the model effectively explore the data space during the data generation, guiding it toward better optimization paths. In contrast, the random data selection strategy showed steeper initial declines in loss but quickly plateaued or fluctuated, particularly in the later stages of training. This points to overfitting caused by repeated exposure to similar samples. These findings demonstrate that the proposed method, generating and selecting diverse data through observing the target model, enables stable training and reduces the risk of overfitting.

D.5 Influence of Different LLMs

We employ different LLMs during augmented data generation to verify the adaptability of the proposed method, as presented in Table 5. Previous research (Zou et al., 2024) points out that different LLMs tend to generate data reflecting their inherent preferences, resulting in data biases. From the results, we can observe that the proposed method shows strong robustness, as using different LLMs does not result in significant performance variations. Since the proposed method does not explicitly constrain the data to align with human-written content, MAUVE* scores exhibit relatively larger changes. However, the scores still surpass those of other DA methods.

D.6 Computational Cost of ODDA

We evaluate OODA's computational cost and efficiency across its components. For LLM APIs, we utilize Python's *asyncio* library for concurrent processing and employ asynchronous interfaces such as *AsyncOpenAI* from LLM vendors.

Selective Demonstration Filter: SDF performs 300 training steps, with logits recorded every 20 steps, requiring under **10 minutes** (9m42s).

Attribute-Constrained Data Generation: The selection of attribute-constrained demonstrations for seed data completes within **5 seconds**, while data augmentation averaged **0.15 seconds** per instance (subject to network latency). It takes 3m42s.

Overall Diversity Data Selection: We utilize an Intel i7-13700K CPU, where the complete Monte Carlo Tree Search (MCTS) process in 8-shot scenarios typically concluded within approximately **10 minutes** (6m43s).

E Case Study

We show examples of the generated augmented data from different DA methods in Table 6. Based on these data, it is clear that the proposed method delivers significantly more diversity, covering different attributes in the sentences, like writing styles, tenses, tones, emotion, and so on. This increased variability underscores the effectiveness of the proposed attribute-constrained data generation and overall diversity data selection module, ultimately enhancing model performance.

Demonstrations

You will be given a triplet of head entity, tail entity, and relation. The following will present several examples from different perspectives. Please refer to them when generating content in order to generate diverse content.

** **Relation Type** **

When a type of relation is given, sentences can be generated like:

1. Relation: <relation>, Sentence: <sentence>.

...

** **Head Entity and Tail Entity** **

When the head entity and the tail entity are given, the length of the dependency path between them in the sentence should be diverse, for example:

1. Head Entity: <head entity>, Tail Entity: <tail entity>, Dependency Path: <dependency path>, Sentence: <sentence>.

...

** **Diverse Semantic** **

The generated sentences should contain different semantic information, for example:

1. Triplet: (<head entity>, <tail entity>, <relation type>), Sentence: <sentence>.

...

** **Diverse Syntax** **

The generated sentences should contain different syntax structures, for example:

1. Triplet: (<head entity>, <tail entity>, <relation type>), Sentence: <sentence>.

...

Task

Your task is to generate *diverse* sentences based on the given triple of head entity, tail entity, and relation. The sentences need to *directly* include the head entity and the tail entity, and there is a given relation between them.

Head Entity: <target head entity>, Tail Entity: <target tail entity>, Relation: <target relation type>

Attributes

To ensure the diversity of sentences, you need to consider the following requirements.

1. Have different writing styles, use simple or complex sentence patterns, and adopt casual, professional, academic, or humorous language contexts.
2. Use different words that can clarify the given relation.
3. Vary the length of sentences and adjust the positions of different parts of the sentences, such as putting adverbial phrases at the beginning, changing the order of the subject and the predicate, etc.
4. Have different voices and perspectives, such as the active and passive voices, the first-person and third-person perspectives, provided that the context allows.
5. If the relationship permits, use the past, present, and future tenses.
6. Have different tones and emotional connotations, such as neutral, enthusiastic, authoritative, or tentative tones.

Output

Note: You must generate diverse sentences.

Generate <number> samples for the relation '<target relation type>', head entity '<target head entity>', and tail entity '<target tail entity>'.

Table 4: The prompt used for attribute-constrained data generation for reference.

LLM	TACRED				ReTACRED				SemEval			
	F1	Self-BLEU	APS	MAUVE*	F1	Self-BLEU	APS	MAUVE*	F1	Self-BLEU	APS	MAUVE*
DeepSeek-v3	75.78	0.4322	0.1189	0.7774	76.78	0.4164	0.1258	0.7532	67.57	0.2406	0.1024	0.9412
GPT-4o	75.80	0.4305	0.1173	0.7528	76.75	0.4161	0.1252	0.7351	67.73	0.2308	0.1076	0.9163
GPT-4o-mini	75.34	0.4386	0.1257	0.7397	76.52	0.4240	0.1338	0.7205	67.16	0.2538	0.1116	0.8925

Table 5: The impact of using different LLMs for data generation on RE performance, data diversity, and quality.

Original Seed Data:

Ferrara said he was innocent of Limoli’s slaying, but he pleaded guilty in 1992 to murder, along with racketeering charges, under a deal that sent him to prison for 22 years, rather than go to trial and risk a conviction that could lead to life in prison.

Triplet: (he, per:charges, murder)

ConsistRE:

1. Witnesses testified against him, leading to the filing of murder charges by the district attorney.
2. He was arrested and now faces murder charges after the victim’s body was discovered last week.
3. He was extradited to face murder charges in the state where the crime occurred.
4. He confessed to the crime, resulting in immediate murder charges being filed.
5. He was acquitted of the murder charges after new evidence proved his alibi.
6. The police arrested him on murder charges after a lengthy manhunt.
7. The court upheld the murder charges against him, rejecting his appeal for dismissal.
8. The defense argued that the murder charges were based on circumstantial evidence.

UnleashLLMRE:

1. He was questioned in connection with the murder case.
2. He was suspected of murder and the investigation was ongoing.
3. He was the prime suspect in the murder investigation.
4. The jury found him guilty, and he was sentenced to life in prison on murder charges.
5. He was believed to have had an accomplice in the murder.
6. He was being held on suspicion of murder.
7. He was known to have a history of violence, including murder.
8. Despite his claims of innocence, he was formally charged with murder and other related offenses.

ODDA (Ours):

1. The court has decided to proceed with murder charges against him.
2. “I didn’t do it," he insisted when questioned about the murder charges.
3. “He is being charged with murder," the detective stated during the press conference.
4. The judge dismissed the murder charges against him due to lack of evidence.
5. Did you hear that he is being accused of murder?
6. Witnesses testified against him, leading to the filing of murder charges by the district attorney.
7. He was charged with murder, along with several other counts, in connection with the case.
8. Authorities allege that he committed the murder last year in cold blood.

Table 6: Examples of the generated augmented data from different LLM-based DA methods.