ReCoT-NER: Enhancing Zero-Shot Named Entity Recognition through Chain-of-Thought Prompting and Recall-Oriented Loss Optimization

Dabin Fu, Fanghong Zhang*
National Center for Applied Mathematics in Chongqing
Chongqing Normal University
Chongqing, China
fanghong.zhang@cqnu.edu.cn

Abstract

Named Entity Recognition (NER) plays a fundamental role in information extraction and domain knowledge construction. However, in specialized domains such as wind power fault diagnosis, the scarcity of labeled data makes supervised approaches impractical. Zero-shot NER provides a promising alternative but still struggles with incomplete entity detection and unstable generation boundaries. To address these challenges, we propose ReCoT-NER, a reasoning-enhanced generative framework that integrates Chain-of-Thought (CoT) prompting and recall-oriented loss optimization. The proposed CoT instruction design explicitly decomposes NER into two reasoning stages: entity span detection and entity type classification. This enables the model to follow a structured inference process. In addition, we introduce a recall-oriented loss function that reweights entity and non-entity tokens to mitigate false negatives, encouraging more inclusive entity coverage. Experiments on CrossNER, MIT, and a newly constructed wind-power NER dataset demonstrate that ReCoT-NER consistently improves recall and overall F1 performance across both general and industrial domains. Notably, ReCoT-NER achieves competitive results with just a 77M-parameter model, making it well-suited for low-resource zero-shot settings.

1 Introduction

Named Entity Recognition (NER) plays a fundamental role in natural language processing (NLP) applications such as information extraction, question answering, and knowledge graph construction(Nasar et al., 2021; Banerjee et al., 2021; Thukral et al., 2023). By identifying entities and their semantic categories, NER systems enable structured understanding of unstructured text(Nadeau and Sekine, 2007). In domain-specific contexts such as wind power fault diagnosis, accurate entity extraction can greatly support intelligent maintenance and operational decision-making. However, conventional supervised NER methods rely heavily on large-scale annotated datasets, which are costly and time-consuming to construct, particularly for specialized technical domains(Li et al., 2022). To alleviate the scarcity of annotated data and improve cross-domain adaptability, recent research has explored generative language models and instruction-tuning paradigms for NER (Wang et al., 2023; Zamai et al., 2024; Zaratiana et al., 2024). These approaches reformulate entity recognition as a text-generation task and leverage instruction tuning and prompt engineering to enhance task adaptability and cross-domain generalization (Wang et al., 2023; Zhou et al., 2023). Subsequent work has introduced targeted techniques, including explicit modeling of negative instances during generation and systematic prompt design, to reduce boundary errors and improve label consistency in cross-domain scenarios (Ding et al., 2024). Despite this progress, existing zero-shot NER systems still face limitations in complex, domain-specific settings: they often produce incomplete extractions, suffer from low recall, and exhibit unstable performance when recognizing implicit entities.

To address these challenges, we propose **ReCoT-NER**, a reasoning-enhanced generative framework that integrates Chain-of-Thought (CoT) prompting with a recall-oriented loss function. The CoT prompting strategy guides the model to decompose the recognition process into two interpretable steps: entity boundary detection and entity type classification, thereby improving semantic transparency and interpretability. In addition, we introduce a modified loss function that prioritizes recall by assigning higher weights to false-negative errors, reducing the likelihood of missing potential entities during generation.

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 4645–4656

We further construct a wind-power domain NER dataset to evaluate zero-shot generalization under domain shift conditions.

The main contributions of this paper are summarized as follows:

- We propose ReCoT-NER, a novel zero-shot NER framework that explicitly models stepwise reasoning through CoT prompting.
- We design a Recall-Oriented Loss (RCLoss) function to improve the detection of challenging or domain-specific entities, particularly those prone to being missed.
- We build a domain-specific NER dataset for wind power fault diagnosis and demonstrate the effectiveness of our approach across both general and industrial corpora. Extensive experiments on seven benchmark domains from CrossNER and MIT show that our method achieves state-of-the-art (SoTA) performance, with the 77M-parameter Flan-T5-small model outperforming prior methods by approximately 2 F1 points.

2 Related Work

2.1 Zero-shot NER

The rapid advancement of large language models (LLMs) has driven a paradigm shift in NER toward more generalizable and instruction-driven frameworks(Ding et al., 2024; Wang et al., 2023). Unlike conventional sequence-labeling methods(Akbik et al., 2018; Chiu and Nichols, 2016; Devlin et al., 2019; Huang et al., 2015), zero-shot NER reframes entity recognition as a reasoning or generation task, allowing models to infer entity boundaries and types based on natural-language descriptions without additional annotated data.

Early studies such as InstructUIE(Wang et al., 2023) were among the first to unify information extraction tasks into a text-to-text formulation, where instruction tuning enhanced the models capability to generalize across multiple tasks and domains. Following this idea, UniversalNER(Zhou et al., 2023) distilled large-scale annotated data synthesized by ChatGPT into LLaMA, yielding strong cross-domain generalization ability. Subsequently, GoLLIE(Sainz et al., 2023) introduced detailed guideline-style prompts that improved the models understanding of complex label definitions and contextual constraints, resulting in notable zero-shot performance gains. Building upon the generative paradigm, GNER(Ding et al., 2024) re-examined the role of negative instances in generative NER and explicitly modeled non-entity labels under a BIO-style framework, significantly enhancing boundary recognition accuracy. SLIMER(Zamai et al., 2024) further incorporated concise entity-type descriptions and task-specific instructions into prompts, enabling the model to better recognize unseen categories. More recently, ZeroNER(Cocchieri et al., 2025) proposed a distillation framework driven by type-aware descriptions, transferring the knowledge of large language models into compact encoder-based models and achieving remarkable performance under strict zero-shot settings. These studies demonstrate that instruction tuning and generative paradigms have shown remarkable advantages in alleviating data scarcity and enhancing cross-domain generalization. However, existing zero-shot NER methods still suffer from low entity recall and high error rates in entity type classification.

2.2 Chain-of-Thought for NER

Recent studies have shown that LLMs can benefit from reasoning-oriented prompting strategies, among which CoT prompting has attracted increasing attention(Wei et al., 2022; Liu et al., 2024). By decomposing complex reasoning into a sequence of interpretable intermediate steps, CoT enables models to generate outputs that are not only accurate but also logically coherent. This paradigm has been successfully applied to various reasoning-intensive tasks, such as mathematical problem solving, commonsense reasoning, and multi-hop question answering(Li et al., 2025; Sun et al., 2025; Mitra et al., 2024; Suzgun et al., 2023).

However, applying CoT prompting to NER remains relatively unexplored. Most existing zero-shot NER approaches rely on concise, task-oriented prompt designs that directly request final entity predictions. Such prompts implicitly induce a single-step generation paradigm, in which entity boundary detection and type classification are performed simultaneously without explicit separation. This joint decision process increases the cognitive burden on the model and often results in incomplete entity spans or incorrect type assignments. Recent studies on structured prompting and task

decomposition suggest that decomposing complex decisions into step-wise reasoning can help mitigate these difficulties. Motivated by this observation, We introduce CoT prompting to explicitly decompose the NER process into two steps: entity span detection and entity type classification.

Specifically, CoT guides the model to perform explicit, interpretable reasoning by first identifying potential entity mentions and then inferring their semantic types based on contextual evidence, ultimately producing the final NER predictions. This step-wise reasoning process not only improves entity recall by encouraging the model to capture implicit or boundary-ambiguous mentions, but also facilitates more accurate type classification through enhanced contextual understanding.

3 Method

3.1 Task Definition

NER aims to identify and classify entity mentions in text into predefined semantic categories such as person, organization, or location. Formally, given an input sentence

$$X = \{x_1, x_2, \ldots, x_n\} \tag{1}$$

where x_i denotes the i-th token, the objective is to assign each token a label

$$Y = \{y_1, y_2, \ldots, y_n\} \tag{2}$$

with $y_i \in \mathcal{L}$, where $\mathcal{L}$ represents the set of entity tags following the BIO scheme.

In the generative formulation adopted by recent LLM-based frameworks, the NER task is reformulated as a conditional text generation problem. Given an instruction-style prompt $P(X)$ that describes the extraction task, the model is trained to generate a corresponding labeled sequence $\hat{Y}$:

$$\hat{Y} = \text{Decoder}(P(X), X) \tag{3}$$

where $\hat{Y}$ contains both the original tokens and their associated entity tags in textual form.

This text-to-text formulation enables a unified generation-based framework, allowing the model to generalize across domains and label sets via natural-language instructions.

3.2 Framework Overview

The overall framework of ReCoT-NER is illustrated in Fig. 1.

3.3 Instruction Tuning

Instruction tuning aims to enhance the adaptability of language models by fine-tuning them on diverse natural-language instructions(Wei et al., 2021; Ouyang et al., 2022). Instead of relying on task-specific classification heads or handcrafted decoding rules, instruction tuning reformulates the task as a unified text-to-text paradigm and trains the model to follow natural-language prompts(Raffel et al., 2020). This paradigm has demonstrated strong cross-task and cross-domain generalization, and has been widely adopted in recent generative NER frameworks.

In this study, we adopt a generative instruction-tuning formulation consistent with the task definition in Section 3.1. Given an input sequence X and its corresponding instruction-style prompt $P(X)$, the model is trained to generate a serialized labeled sequence $\hat{Y}$:

$$\hat{Y}^* = \arg\max_{\hat{Y}} \; p(\hat{Y} \mid [P(X); X]; \theta) \tag{4}$$

where θ denotes the model parameters. The prompt $P(X)$ explicitly specifies the extraction task and constrains the output format, guiding the model to produce token–label pairs that follow the BIO tagging scheme.

During task-adaptive training, we utilize the Pile-NER(Zhou et al., 2023) dataset as the instruction-tuning corpus. This dataset spans multiple domains and entity types, enabling the model to acquire general-purpose entity extraction capabilities prior to zero-shot evaluation. Following standard practice in generative NER, the training objective minimizes the token-level negative log-likelihood over the generated sequence:

$$\mathcal{L}_{\text{gen}} = -\sum_{t=1}^{T} \log p(\hat{y}_t \mid \hat{y}_{<t}, P(X), X; \theta) \tag{5}$$

where T denotes the length of the generated output sequence, $\hat{y}_t$ is the token generated at decoding step t, and $\hat{y}_{<t} = (\hat{y}_1, \ldots, \hat{y}_{t-1})$ denotes the sequence of previously generated tokens.

Through instruction tuning, the model learns to align natural-language task descriptions with structured entity extraction behaviors, effectively bridging language generation and sequence labeling(Xia et al., 2023). This design allows the model to generalize to unseen domains and entity types without requiring any labeled data from the target task.

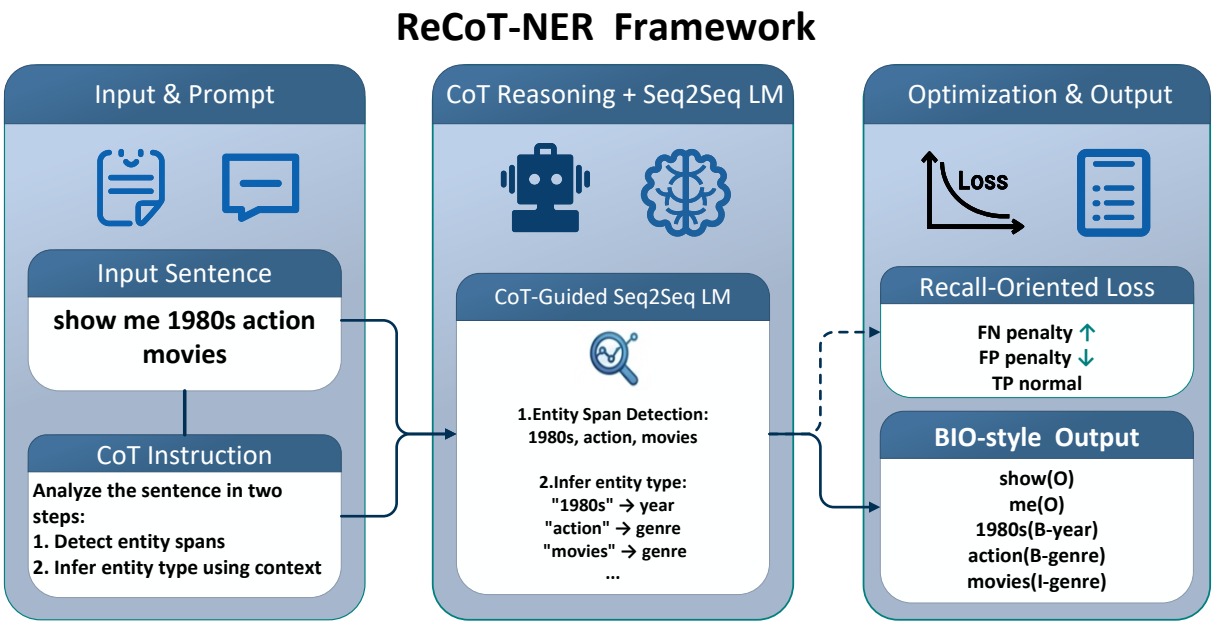

Figure 1: Overview of the ReCoT-NER framework. Given an input sentence, a CoT instruction explicitly decomposes the NER task into two stages: entity span detection and entity type classification. The prompted input is processed by a CoT-guided sequence-to-sequence language model, which performs implicit reasoning during autoregressive decoding to generate BIO-style entity predictions. During training, the model is optimized with a recall-oriented loss that selectively reweights token-level learning signals to emphasize entity boundary , thereby improving entity recall in zero-shot generative NER.

3.4 CoT Instruction Design

To enhance the reasoning capability of generative NER, we introduce a CoT-based instruction design(Longpre et al., 2023). Existing generative NER methods typically adopt concise, task-oriented prompts that directly request final entity predictions, without explicitly guiding the model through entity span detection and type inference. Such prompt designs provide limited semantic structure, making it difficult to reliably infer implicit entities or disambiguate highly confusable entity types.

Motivated by the inherent two-stage nature of the NER task, we design CoT-guided instructions that explicitly decompose entity recognition into two sequential reasoning steps: (i) entity span detection and (ii) entity type classification. As shown in Fig. 2, the proposed prompt makes this reasoning structure explicit at the instruction level, guiding the model to first identify candidate entity spans based on contextual cues and then assign appropriate semantic labels.

Each input instance is reformulated into a structured instruction that specifies the task objective, output format, and reasoning steps. This design encourages the model to follow a coherent reasoning path during decoding, effectively decoupling span detection from type classification. The intermediate reasoning steps act as implicit supervision, helping the model learn interpretable extraction behaviors.

To ensure instruction consistency across different domains, we adopt a unified prompt format while varying the entity type list according to the target dataset. This strategy enables stable generation and robust adaptation in zero-shot settings.

3.5 Improved Loss Function

Although instruction tuning enables language models to generalize across domains, generative NER frameworks often suffer from low entity recall and incomplete span generation in zero-shot settings. This issue is closely related to the imbalance in the generative decoding process: the output sequence typically contains a large number of non-entity tokens, while entity-related tokens occur sparsely(Lin et al., 2017). Under the standard cross-entropy (CE) objective, all generated tokens are penalized uniformly, causing the optimization process to be dominated by frequent non-entity tokens. As a result, the training signal associated with entity-related tokens becomes relatively weak, leading the model to favor conservative predictions and miss potential entities.

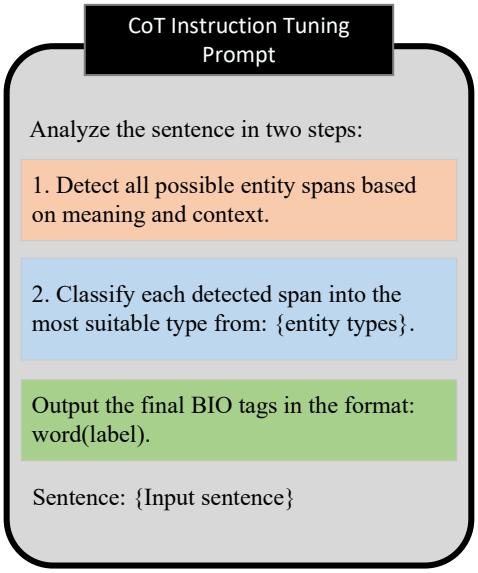

Figure 2: The CoT instruction tuning prompt used in ReCoT-NER.

A natural idea to improve recall is to directly amplify the penalties of entity tokens. However, in zero-shot settings, the entity types appearing in the target domain are unknown during task-adaptive training. Introducing type-specific weighting would therefore violate the zero-shot assumption or require access to unavailable supervision. This constraint motivates a more structure-aware solution that does not rely on entity category information.

To this end, we propose a recall-oriented loss function, termed **RCLoss**, which leverages the intrinsic structure of the BIO tagging scheme. Instead of emphasizing specific entity types, RCLoss focuses on strengthening the supervision of entity boundary signals. The key observation is that the `B` token serves as the unique indicator of an entity span start. In an autoregressive generation framework, increasing the confidence of `B` tokens naturally facilitates the subsequent generation of subsequent entity-type tokens and `I` tokens, thereby improving complete span recall. A supporting analysis is provided in Appendix C.

Formally, given the predicted probability p_t assigned by the decoder to the generated token $\hat{y}_t$ at decoding step t, the token-level loss is defined as:

$$\mathcal{L}_{\text{recall}} = w_t \cdot (-\log p_t),$$
$$w_t = \begin{cases} \alpha(1-p_t), & \hat{y}_t = \texttt{B token}, \\ \beta(1-p_t), & \hat{y}_t \neq \texttt{B token}. \end{cases} \quad (6)$$

where α and β are weighting coefficients satisfying $\alpha > \beta > 0$. This design selectively amplifies the learning signal of entity-start predictions, while maintaining moderate penalties for other tokens. The confidence-aware factor $(1-p_t)$ further encourages the model to focus on low-confidence target predictions, allowing potentially missed entity boundaries to receive stronger supervision without excessively penalizing confident outputs.

4 Experiments

4.1 Experimental Settings

4.1.1 Datasets

The Pile-NER(Zhou et al., 2023) dataset was employed as the task-adaptive instruction-tuning dataset, with CrossNER(Liu et al., 2021), MIT(Liu et al., 2013), and Wind-Power NER used for zero-shot evaluation.

Pile-NER.(Zhou et al., 2023) Following Zhou et al., we adopt Pile-NER as the task-adaptive training corpus. This dataset is derived from the Pile corpus, containing approximately 240K entities across 13K distinct entity categories. The data were automatically annotated using ChatGPT to produce high-quality entity spans in an open-domain context. It covers diverse domains such as news, medicine, law, and entertainment, providing rich linguistic and entity-type variations. Training on Pile-NER enables the model to acquire general extraction abilities before being evaluated on unseen domains.

CrossNER and MIT. CrossNER (Liu et al., 2021) and MIT (Liu et al., 2013) are used as zero-shot evaluation benchmarks. CrossNER covers five domains, including literature, AI, politics, science, and music. Following the settings of GNER and UniNER, we remove the "else" entity type from CrossNER to maintain a consistent label schema and avoid type ambiguity. The MIT dataset consists of two subsets, namely MIT-Movie and MIT-Restaurant. For both benchmarks, we use the standard data splits adopted in prior work.

Wind-Power NER. To further assess zero-shot generalization in specialized technical domains, we construct wind-power NER, a domain-specific NER dataset focused on the wind energy sector. The dataset is curated from real-world operational documents, including wind farm maintenance reports, technical manuals, and peer-reviewed research articles in wind energy engineering. The

dataset includes named entities referring to turbine components, operational conditions and failure modes, annotated in BIO format. Unlike open-domain benchmarks, wind-power NER reflects the lexical specificity, contextual ambiguity, and terminological complexity characteristic of industrial technical writing. As a challenging zero-shot evaluation benchmark, it enables us to rigorously test whether ReCoT-NER can generalize to unseen, highly specialized terminology. Detailed dataset statistics and construction details are provided in Appendix D.

4.1.2 Baselines

The following models are used as baselines for zero-shot NER:

InstructUIE(Wang et al., 2023) was among the first to unify information extraction tasks under a text-to-text formulation, where instruction tuning enables the model to generalize across diverse tasks and domains.

UniNER(Zhou et al., 2023) extended this paradigm by distilling large-scale synthetic annotations generated by ChatGPT into the LLaMA model, achieving strong cross-domain generalization and robustness to unseen entity types.

GoLLIE(Sainz et al., 2023) introduced detailed guideline-style prompts that improved the models comprehension of complex label definitions and contextual constraints, yielding remarkable zero-shot performance gains.

GNER(Ding et al., 2024) re-examined the use of negative instances in generative NER, explicitly modeling non-entity tokens within a BIO-style framework to significantly enhance entity boundary recognition.

4.1.3 Metrics

We evaluate the model performance using the micro F1 score, which is a widely adopted metric in NER tasks. The F1 score is defined as the harmonic mean of Precision and Recall, formulated as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{8}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

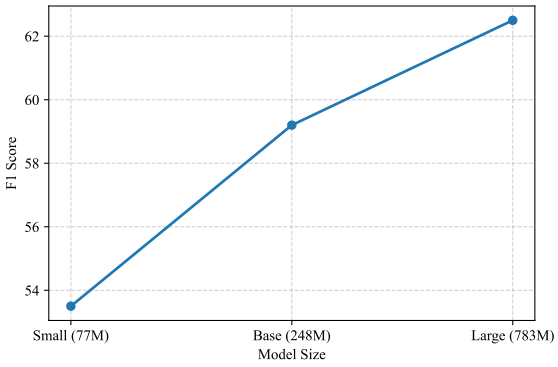

Figure 3: Effect of model scaling on zero-shot F1 performance across different backbone sizes.

where TP denotes the number of correctly recognized entity spans, FP represents the number of incorrectly predicted entities, and FN indicates the number of missed entities.

4.2 Results

4.2.1 Zero-shot Evaluation

Table 1 summarizes the zero-shot performance across the benchmarks: CrossNER and MIT. Across all backbone sizes, the proposed ReCoT-NER consistently outperforms the reproduced GNER baselines, demonstrating the effectiveness of integrating CoT reasoning with a recall-oriented loss. While ReCoT-NER does not always achieve the best score in every individual domain, it obtains the highest average F1 at each model scale, reflecting strong overall robustness. Notably, even the compact Flan-T5-small model achieves an average F1 of 53.5, surpassing several much larger instruction-tuned systems such as InstructUIE-11B and UniNER-7B. This highlights the efficiency of our design in small-model settings.

In addition, as illustrated in Fig. 3, scaling the backbone from 77M to 783M parameters yields a clear and stable improvement in average F1, increasing from 53.5 to 62.5. This monotonic trend confirms that the proposed CoT-guided reasoning and recall-oriented optimization remain effective as model capacity grows.

To assess the generalization ability of the framework under more challenging distribution shifts, we further evaluate ReCoT-NER on the Wind-Power NER dataset. As shown in Table 2, ReCoT-NER outperforms GNER-T5 across all configurations. The Flan-T5-small model achieves an F1 of 41.71, and performance improves with model

Method	Backbone	AI	Literature	Music	Politics	Science	Movie	Restaurant	Avg
ChatGPT	–	52.4	39.8	66.6	68.5	67.0	5.3	32.8	47.5
InstructUIE	flan-t5-xxl (11B)	48.4	48.8	54.4	49.9	49.4	**63.0**	21.0	47.8
GoLLIE-7B	Code-LLaMA-7B	60.3	**67.1**	64.5	60.8	60.5	**63.0**	43.4	59.9
GoLLIE-13B	Code-LLaMA-13B	**63.8**	60.1	68.5	56.2	61.5	62.5	**49.8**	60.3
UniNER-7B	LLaMA-7B	53.5	59.4	65.0	60.8	61.1	42.4	31.7	53.4
UniNER-13B	LLaMA-13B	54.2	60.9	64.5	61.4	63.5	48.7	36.2	55.6
GNER-T5	flan-t5-small (77M)	51.0	51.9	67.1	58.9	62.1	41.7	28.5	51.6
	flan-t5-base (248M)	56.1	57.6	69.9	60.9	63.7	59.5	40.6	58.3
	flan-t5-large (783M)	60.6	57.9	75.6	62.1	68.7	57.3	45.5	61.1
ReCoT-NER	flan-t5-small (77M)	53.9	57.0	67.5	60.3	63.2	42.7	29.8	53.5
	flan-t5-base (248M)	54.6	59.5	71.7	68.3	64.1	56.1	39.9	59.2
	flan-t5-large (783M)	58.5	58.4	**77.8**	**69.2**	**70.0**	57.0	46.5	**62.5**

Table 1: Zero-shot evaluation results on the benchmark. The best results are highlighted in bold, while the second-best results are underlined. Results for ChatGPT and UniNER are from(Zhou et al., 2023); InstructUIE are from(Wang et al., 2023); GoLLIE are from(Sainz et al., 2023).GNER-T5 denotes our reimplementation under the same training settings as our method.

Method	Backbone	Wind-Power
GNER-T5	flan-t5-small (77M)	38.34
	flan-t5-base (248M)	47.61
	flan-t5-large (783M)	53.08
ReCoT-NER	flan-t5-small (77M)	41.71
	flan-t5-base (248M)	50.10
	flan-t5-large (783M)	61.34

Table 2: Zero-shot evaluation results on the Wind-Power NER dataset, which serves as an out-of-domain testbed to assess cross-domain generalization.

scale, reaching 61.34 with Flan-T5-large, demonstrating superior performance in specialized industrial domains.

Overall, the results across both open-domain and industrial-domain evaluations demonstrate that ReCoT-NER not only improves zero-shot performance on standard benchmarks but also transfers effectively to real-world specialized scenarios.

4.2.2 Ablation Results

To assess the contribution of each component in our framework, we perform ablation studies on the Flan-T5-small backbone, examining CoT and RCLoss. The results in Table 3 show that removing either component leads to a clear decrease in performance, indicating that both are essential to the effectiveness of the proposed approach.

When CoT is removed, the average F1 score decreases from 53.55 to 52.59. This decline demonstrates that CoT-guided reasoning provides valuable structural guidance, allowing the model to separate span detection from type classification more effectively. Without this reasoning process, the model tends to generate incomplete spans or uncertain label assignments, especially in domains with complex contextual dependencies such as Literature and Movie.

Eliminating RCLoss results in an even larger reduction in performance, lowering the average F1 score to 51.32. This weakness stems from the fact that generative NER models naturally under-extract entities in zero-shot settings, and RCLoss directly compensates for this issue by increasing the penalty assigned to low-confidence entity predictions. Without RCLoss, the model more frequently misses valid spans.

Overall, the full model achieves the highest scores across almost all domains, showing that CoT and RCLoss improve orthogonal aspects of the task. CoT enhances structured reasoning and span completeness, whereas RCLoss improves the models ability to recall entity mentions more effectively. Their combination yields more stable and generalizable zero-shot performance.

4.3 Analysis

4.3.1 Impact of CoT Prompting

To assess the impact of CoT prompting, we compare the full ReCoT-NER model with its ablated variant without CoT. Results across four representative domains, namely AI, Literature, Politics, and Movie, are shown in Table 4. Overall, the removal of CoT consistently leads to performance degradation, with varying degrees across domains.

On the AI domain, the recall increases slightly by 0.44 points; however, this marginal gain is accompanied by a notable decrease of 2.57 points

Method	AI	Literature	Music	Politics	Science	Movie	Restaurant	Avg
Full model	**53.96**	**57.09**	**67.55**	**60.38**	63.28	**42.75**	29.85	**53.55**
w/o CoT	52.62	54.17	66.40	58.89	**63.80**	40.21	**32.02**	52.59
w/o RCLoss	51.59	50.47	67.16	57.95	63.36	39.79	28.92	51.32

Table 3: Ablation study of the CoT and the RCLoss on the Flan-T5-small model.

Method	AI (R / P)	Literature (R / P)	Politics (R / P)	Movie (R / P)
Full model	58.24 / 50.26	54.28 / 60.21	61.28 / 59.51	39.38 / 46.74
w/o CoT	58.68 / 47.69	51.86 / 56.70	59.43 / 58.35	36.46 / 44.81
Δ (w/o CoT)	+0.44 / -2.57	-2.42 / -3.51	-1.85 / -1.16	-2.92 / -1.93
w/o RCLoss	54.82 / 48.72	47.93 / 53.31	57.79 / 58.12	34.87 / 46.34
Δ (w/o RCLoss)	-3.42 / -1.54	-6.35 / -6.90	-3.49 / -1.39	-4.51 / -0.40

Table 4: Impact of CoT prompting and effect of the RCLoss. R and P denote recall and precision, respectively.

in precision. This suggests that the models ability to identify entity spans remains essentially unchanged without CoT guidance, while its ability to correctly assign entity labels diminishes, leading to a noticeable drop in precision. In the other three domains, the influence of CoT becomes more evident, with significant drops in both recall and precision.

These results collectively demonstrate that CoT prompting offers consistent benefits for both boundary detection and label assignment. By introducing an intermediate reasoning step, CoT encourages the model to first identify potential entity spans before committing to final label generation. This explicit reasoning structure stabilizes entity extraction, reduces spurious predictions, and enhances the models ability to retain subtle contextual cues, which is particularly important in zero-shot scenarios where no domain-specific supervision is available.

4.3.2 Effect of the Recall-Oriented Loss

The effect of the recall-oriented loss is shown in Table 4. When RCLoss is removed, all domains experience notable drops in recall, confirming that the loss function plays a central role in strengthening the models sensitivity to entity spans. The Literature domain shows the most substantial decline, with recall decreasing from 54.28 to 47.93, a reduction of 6.35 points. This indicates that recall-focused weighting is particularly important in domains containing diverse entity categories and more complex linguistic structures. This pattern highlights that the adaptive weighting mechanism effectively reduces missed detections, which is essential for zero-shot NER scenarios where entity boundaries and types vary widely across domains.

In addition to improving recall, the loss function also yields a mild but consistent improvement in precision. This suggests that encouraging the model to attend more strongly to potential entity tokens does not compromise its discrimination ability; instead, the strengthened span awareness helps the model avoid ambiguous or incomplete predictions, leading to a slight gain in precision as well.

The results show that RCLoss enhances recall, which is the primary design objective, while simultaneously preserving or even improving precision, yielding a more balanced and robust entity detection capability across diverse zero-shot scenarios.

5 Conclusion

This paper presents ReCoT-NER, a generative zero-shot NER framework that integrates CoT prompting with a recall-oriented optimization objective. By explicitly decomposing entity extraction into entity span detection and entity type classification, ReCoT-NER introduces structured reasoning into the generative NER process. The proposed recall-oriented loss further emphasizes low-confidence entity predictions, effectively reducing false negatives and improving span completeness without compromising classification reliability.

Extensive experiments on the CrossNER and MIT benchmarks demonstrate that ReCoT-NER consistently enhances zero-shot performance across diverse domains and model scales, with particularly pronounced gains for compact models. Furthermore, evaluation on wind-power NER shows that ReCoT-NER maintains robust general-

ization to unseen, highly specialized terminology. These results further demonstrate the effectiveness of structured reasoning and recall-aware optimization in challenging zero-shot scenarios.

Limitations

Although proposed approach achieves competitive results in zero-shot NER. This work is subject to several limitations that suggest directions for further investigation.

First, the incorporation of CoT is confined to instruction-level reasoning guidance. Although the prompt design explicitly encourages a two-stage reasoning process, the reasoning itself remains implicit during generation, without explicit modeling or supervision of intermediate reasoning steps. This design may limit the models ability to fully capture complex semantic dependencies in highly specialized or technical texts.

Second, the proposed recall-oriented loss function is specifically designed to alleviate false negatives by strengthening the learning signal for entity boundary-related tokens. While this strategy effectively improves entity recall, it does not explicitly target fine-grained entity type discrimination.

Finally, due to hardware constraints, empirical evaluation is conducted only on language models with relatively small parameter sizes. Although the proposed method demonstrates strong effectiveness under this setting, its behavior and scalability on larger instruction-tuned models remain to be explored. Future work will extend the evaluation to larger-scale models to further assess the generality of the proposed approach.

Acknowledgments

This work was supported by the Science and Technology Innovation Key R&D Program of Chongqing (CSTB2025TAD-STX0025) and the foundation of Chongqing Normal University under Grant 25XLB030. We also thank the National Advanced Computing Laboratory in Taiyuan and Chongqing for providing computational resources.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Pratyay Banerjee, Kuntal Kumar Pal, Murthy Devarakonda, and Chitta Baral. 2021. Biomedical named entity recognition via knowledge guidance and question answering. *ACM Trans. Comput. Healthcare*, 2(4).

Jason P.C. Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Alessio Cocchieri, Marcos Martínez Galindo, Giacomo Frisoni, Gianluca Moro, Claudio Sartori, and Giuseppe Tagliavini. 2025. ZeroNER: Fueling zero-shot named entity recognition via entity type descriptions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15594–15616, Vienna, Austria. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yuyang Ding, Juntao Li, Pinzheng Wang, Zecheng Tang, Yan Bowen, and Min Zhang. 2024. Rethinking negative instances for generative named entity recognition. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3461–3475, Bangkok, Thailand. Association for Computational Linguistics.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2022. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Ronghan Li, Dongdong Li, Haowen Yang, Xiaoxi Liu, Haoxiang Jin, RongCheng Pu, and Qiguang Miao. 2025. Recot: Relation-enhanced chains-of-thoughts for knowledge-intensive multi-hop questions answering. *Neurocomputing*, 637:129903.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Jingjing Liu, Panupong Pasupat, Scott Cyphers, and Jim Glass. 2013. Asgard: A portable architecture for multilingual dialogue systems. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8386–8390.

Yanming Liu, Xinyue Peng, Tianyu Du, Jianwei Yin, Weihao Liu, and Xuhong Zhang. 2024. ERA-CoT:

Improving chain-of-thought through entity relationship analysis. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8780–8794, Bangkok, Thailand. Association for Computational Linguistics.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2021. Crossner: Evaluating cross-domain named entity recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13452–13460.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 22631–22648. PMLR.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2024. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14431.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik. 2021. Named entity recognition and relation extraction: State-of-the-art. *ACM Comput. Surv.*, 54(1).

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. 2023. Gollie: Annotation guidelines improve zero-shot information-extraction. *arXiv preprint arXiv:2310.03668*.

Jun Sun, Yiteng Pan, and Xiaohu Yan. 2025. Improving intermediate reasoning in zero-shot chain-of-thought for large language models with filter supervisor-self correction. *Neurocomputing*, 620:129219.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.

Anjali Thukral, Shivani Dhiman, Ravi Meher, and Punam Bedi. 2023. Knowledge graph enrichment from clinical narratives using nlp, ner, and biomedical ontologies for healthcare applications. *International Journal of Information Technology*, 15(1):53–65.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, and 1 others. 2023. Instructuie: Multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yu Xia, Yongwei Zhao, Wenhao Wu, and Sujian Li. 2023. Debiasing generative named entity recognition by calibrating sequence likelihood. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1137–1148, Toronto, Canada. Association for Computational Linguistics.

Andrew Zamai, Andrea Zugarini, Leonardo Rigutini, Marco Ernandes, and Marco Maggini. 2024. Show less, instruct more: Enriching prompts with definitions and guidelines for zero-shot ner. *arXiv preprint arXiv:2407.01272*.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico. Association for Computational Linguistics.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2023. Universalner: Targeted distillation from large language models for open named entity recognition. *arXiv preprint arXiv:2308.03279*.

A Implementation Details

All experiments are implemented using the PyTorch 2.4.0 framework with the DeepSpeed library for distributed and mixed-precision training. The Flan-T5 series, including Flan-T5-small, Flan-T5-base, and Flan-T5-large, was evaluated to assess the effect of model scale, with Flan-T5-base serving as the primary model.

During the task adaptation stage, the model is fine-tuned on the Pile-NER dataset using the instruction-tuning paradigm. We set the maximum input and output sequence length to 640 tokens, a batch size of 16. To account for differences in model capacity and training stability, the number of training epochs varies across model sizes: Flan-T5-small and Flan-T5-base are trained for 20 epochs, while Flan-T5-large is trained for 6 epochs. The AdamW(Loshchilov and Hutter, 2018) optimizer is used with a fixed learning rate of 5×10^{-5}, and no weight decay. A constant learning rate scheduler is applied with no warm-up steps. The proposed recall-oriented loss uses $\alpha = 1.2$ for B tokens and $\beta = 1.0$ for all other tokens. The loss function is optimized using the proposed recall-oriented adaptive objective. We employ bf16 mixed-precision training for efficiency and stability. Model evaluation is performed at the end of each epoch, selecting the best checkpoint based on the highest average F1 score on the validation set. All experiments are conducted on a workstation equipped with two NVIDIA RTX 4090 GPUs, each with 24 GB of memory.

B RCLoss Hyperparameter Sensitivity Analysis

To further examine the effect of the key hyperparameter in RCLoss, we conduct a sensitivity analysis on the Flan-T5-small model. In RCLoss, α controls the weight of B tokens, while β controls the weight of all other tokens. Since our method mainly aims to strengthen the learning signal on entity-start positions, we focus on the influence of α in this subsection.

Specifically, we fix $\beta = 1.0$ and vary α from 1.0 to 2.0 with a step size of 0.1. We report Average Precision, Average Recall, and Average F1 on the general-domain test sets, including all CrossNER subdomains as well as MIT-Movie and MIT-Restaurant. The results are shown in Fig. 4.

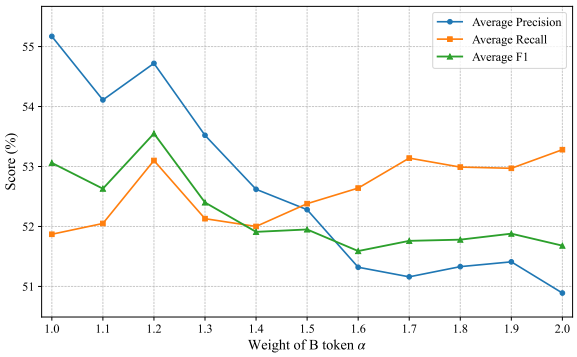

Figure 4: Effect of the B-token weight α on Average Precision, Average Recall, and Average F1 of Flan-T5-small, with $\beta = 1.0$ fixed.

As α increases, Average Recall generally shows an upward trend, whereas Average Precision gradually decreases. This observation is consistent with the design motivation of RCLoss: assigning a larger weight to B tokens makes the model more sensitive to entity-start positions, which helps recover more entities, but also increases the risk of predicting non-entity tokens as entity starts, thereby introducing more false positives.

From the perspective of the overall trade-off, Average F1 first rises and then declines. When α increases from 1.0 to 1.2, Average F1 improves from 53.06 to 53.55, indicating that a moderate increase in the B token weight can improve recall while still maintaining relatively stable precision. However, when α becomes larger, the gain in recall is gradually offset by a more obvious drop in precision, which eventually leads to lower F1.

Overall, these results indicate that model performance is moderately sensitive to the choice of α. A moderate range of α yields relatively stable performance, whereas overly small or overly large values lead to noticeable degradation. This suggests that the recall-oriented enhancement on entity-start positions should be kept within a proper range. According to Fig. 4, the best Average F1 is achieved at $\alpha = 1.2$. Therefore, we set $\alpha = 1.2$ and $\beta = 1.0$ as the default hyperparameters of RCLoss in all experiments.

C Tokenization Analysis

In this section, we analyze the tokenization outputs of the Flan-T5 variants and BERT on a BIO-

Model	Input Text	Tokenization Output
Flan-T5-Small	Regarding(O) GERD(B-group) patients(B-group) without(O) globus(O)	['Regard', 'ing', '(', 'O', ')', '', 'GER', 'D', '(', 'B', '-', 'group', ')', 'patients', '(', 'B', '-', 'group', ')', 'without', '(', 'O', ')', '', 'glob', 'us', '(', 'O', ')']
Flan-T5-Base	Regarding(O) GERD(B-group) patients(B-group) without(O) globus(O)	['Regard', 'ing', '(', 'O', ')', '', 'GER', 'D', '(', 'B', '-', 'group', ')', 'patients', '(', 'B', '-', 'group', ')', 'without', '(', 'O', ')', '', 'glob', 'us', '(', 'O', ')']
Flan-T5-Large	Regarding(O) GERD(B-group) patients(B-group) without(O) globus(O)	['Regard', 'ing', '(', 'O', ')', '', 'GER', 'D', '(', 'B', '-', 'group', ')', 'patients', '(', 'B', '-', 'group', ')', 'without', '(', 'O', ')', '', 'glob', 'us', '(', 'O', ')']
BERT (bert-base-cased)	Regarding(O) GERD(B-group) patients(B-group) without(O) globus(O)	['Regarding', '(', 'O', ')', 'GE', '##RD', '(', 'B', '-', 'group', ')', 'patients', '(', 'B', '-', 'group', ')', 'without', '(', 'O', ')', 'g', '##lo', '##bus', '(', 'O', ')']

Table 5: Tokenization results of a BIO-annotated sentence using different pretrained language models.

Model	Run 1	Run 2	Run 3	Average F1	Std
Flan-T5-Small	0.5381	0.5358	0.5328	0.5356	0.0027
Flan-T5-Base	0.5883	0.5918	0.5961	0.5921	0.0049
Flan-T5-Large	0.6247	0.6254	0.6261	0.6254	0.0007

Table 6: Average F1 scores over three independent runs. Std denotes sample standard deviation.

annotated sentence. As shown in Table 5, the boundary marker, which denotes the start of an entity span in the BIO tagging format, is consistently represented as an independent token across different tokenizers. This observation supports the feasibility of the proposed loss design, as it enables boundary-aware weighting of token-level training signals by explicitly emphasizing the generation of B tokens during autoregressive decoding.

D Dataset Details

We construct a domain-specific Named Entity Recognition dataset for wind power fault diagnosis, referred to as Wind-Power NER, to evaluate zero-shot generalization in a specialized industrial domain. The dataset is written in English and consists of 1,312 sentences, split into 656 training, 262 development, and 394 test instances, with a total of 2,384 annotated entities. The dataset defines 11 domain-relevant entity types, including equipment, event, environmental condition, equipment status, technology, and related contextual categories. All annotations follow the BIO scheme and are provided in the standard CoNLL format. The dataset is curated from wind power domain texts, including real-world operational documents, technical manuals, and research materials. To protect privacy and ensure research compliance, the collected texts were processed to remove personally identifiable information where necessary. The dataset is used exclusively for zero-shot evaluation, without any task-adaptive training on the target domain.

E Experimental Reproducibility and AI Tool Usage

E.1 Result Reporting.

To ensure the reliability of the experimental results, all evaluations are conducted over three independent runs with different random seeds. The averaged F1 scores and the corresponding standard deviations are reported in Table 6.

As shown in Table 6, the performance remains highly stable across repeated runs for all model sizes. The standard deviations are consistently small, indicating that the proposed framework produces reproducible results and is not sensitive to random initialization.

E.2 Use of AI-Assisted Tools.

We used GitHub Copilot solely as a programming aid to improve code readability and implementation efficiency. No AI tools were used to generate scientific content, including model architectures, loss function designs, experimental protocols, or result analysis. All code and experiments were manually reviewed, validated, and executed by the authors to ensure correctness and reproducibility.