

NE-R1: Enhancing Named Entity Recognition Model via Reinforcement Learning

Meixuan Chen¹ * ‡, Hehan Li²*, Ruizhi Zhao³, Xin Lu², Peizhi Xu²,
Liwei Qian², Meifang Li², Shuanglong Li², Hanmeng Liu²,
Xin Pei², Yanbiao Ma^{4,5,6†}

¹Shenzhen Graduate School, Peking University

²Baidu, Inc., Beijing, China

³Shenzhen University, Shenzhen, China

⁴Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China

⁵Beijing Key Laboratory of Research on Large Models and Intelligent Governance

⁶Engineering Research Center of

Next-Generation Intelligent Search and Recommendation, MOE

chenmeixuan26@stu.pku.edu.cn 2410145008@mails.szu.edu.cn ybma1998@ruc.edu.cn

{lihehan, luxin06, xupeizhi, qianliwei, limeifang, lishuanglong, liuhanmeng}@baidu.com

Abstract

Named Entity Recognition (NER) has achieved substantial progress since the advent of large language models (LLMs). Nevertheless, the recognition of long-tail and domain-specific entities remains challenging due to the deficiency in parametric knowledge. Retrieval-augmented generation (RAG) offers a promising remedy by injecting external knowledge, but it also introduces noise and unnecessary cost when dealing with familiar cases. In this paper, we propose **NE-R1**, a novel framework for adaptive retrieval-augmented NER. We design a “retrieval-on-demand” mechanism for NER. Then we integrate it into models by a two-stage training method: (1) multi-task instruction tuning initialization; (2) end-to-end RL optimization with CoT. To achieve reasonable selection between parameterized and external knowledge, we design a multi-dimensional reward considering both accuracy and retrieval benefit. NE-R1 achieves state-of-the-art performance on various benchmarks, with average improvements of **2.52** F1 points in in-domain evaluation and **1.18** F1 points in zero-shot cross-domain evaluation.

1 Introduction

Named Entity Recognition (NER) is a fundamental task in Natural Language Processing (NLP), aiming to identify entity spans and their types (Lample et al., 2016; Ma and Hovy, 2016). In recent years, large language models (LLMs) have catalyzed the rapid advancement of Generative NER (GenNER) (Seow et al., 2025; Wang et al., 2023;

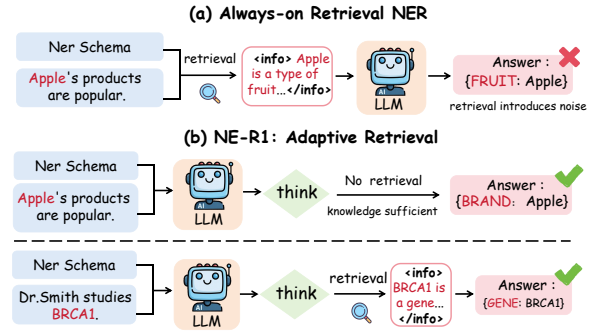


Figure 1: Motivation and overview of NE-R1’s adaptive retrieval mechanism. (a) Always-on retrieval indiscriminately introduces redundant noise and degrades performance for familiar entities. (b) NE-R1 adaptively triggers retrieval only when genuinely necessary.

Zhou et al., 2024), which reformulates NER as a conditional generation problem, and achieved remarkable progress in this area.

However, GenNER still suffers from the parametric knowledge deficiency of LLMs (Lu et al., 2022; Guo et al., 2025b). This leads to hallucinations or omissions when encountering long-tail, domain-specific unfamiliar entities or ambiguous situations (Kang et al., 2025). To address this limitation, Retrieval-Augmented Generation (RAG) (Gao et al., 2023b; Fan et al., 2024) has been introduced to inject external knowledge into NER models (Tan et al., 2023; Zhenwei et al., 2024), and further augmentation has been explored through retrieving similar examples (Li et al., 2024).

Despite the benefits brought by retrieval augmentation, existing methods predominantly adopt a “full retrieval” strategy—triggering retrieval indiscriminately regardless of input complexity. This

*Equal contribution. † Corresponding author.

‡ Work done during an internship at Baidu, Inc.

paradigm engenders two critical issues: **(1) Noise Introduction**, where mandatory retrieval for familiar cases may inject irrelevant information that interferes with model decisions, as illustrated in Figure 1(a). **(2) Unnecessary Latency Overhead**, where retrieval for high-frequency entities is often redundant. Our experiments on the MIT-Movie dataset reveal that full retrieval incurs a $4.85\times$ increase in inference latency compared to no retrieval, yet yields negligible performance gains. Consequently, it becomes paramount to realize an adaptive “**retrieve-on-demand**” mechanism and seamlessly integrate it into NER models.

In this paper, we propose **NE-R1**, a novel framework for adaptive retrieval-augmented NER. As shown in Figure 1(b), NE-R1 intelligently determines when to trigger retrieval based on input complexity. Inspired by the recent breakthroughs of DeepSeek-R1 (Guo et al., 2025a) in reasoning tasks, we design a two-stage training procedure. The first stage is capability initialization through multi-task instruction tuning, which endows the model with three foundational capabilities: parametric inference, retrieval triggering, and evidence fusion; the second stage is end-to-end RL training, incorporating an explicit Chain-of-Thought mechanism and a multi-dimensional reward that comprehensively estimates prediction correctness and retrieval benefit to jointly enhance overall performance.

Extensive experiments show that NE-R1 outperforms the evaluated baselines, improving the average in-domain F1 score by **2.52** points over the strongest overall baseline, including a **3.30**-point gain on GENIA. In zero-shot cross-domain evaluation, NE-R1 improves performance in four of the five CrossNER domains and achieves the best five-domain average, exceeding the strongest baseline by **1.18** F1 points. On MIT-Movie, NE-R1 also reduces inference latency by approximately **55%** compared with the always-retrieve baseline while achieving a higher F1 score. Our contributions are summarized as follows:

- We propose NE-R1, a novel framework that brings a “retrieve-on-demand” mechanism to NER, allowing models to retrieve only when their own knowledge falls short.
- We design a two-stage training method that first injects foundational capabilities via multi-task instruction tuning, then performs end-to-end RL op-

timization with Chain-of-Thought reasoning and a multi-dimensional reward.

- NE-R1 surpasses existing methods on both in-domain and zero-shot cross-domain evaluations, demonstrating an favorable balance of performance, efficiency, and generalization.

2 Related Work

LLM-based NER. Recent LLM-based NER work primarily follows two main paradigms. The first is In-Context Learning (Dong et al., 2024), which directly invokes LLMs through prompts. Works in this direction focus on instruction optimization (Liu et al., 2023; Wang et al., 2025b; Gonen et al., 2023), providing domain knowledge (Tong et al., 2025; Cocchieri et al., 2025; Huang et al., 2025b), or introducing reflection mechanisms (Ashok and Lipton, 2023; Xie et al., 2023). ICL example selection (Xie et al., 2024; Chen et al., 2025) is also explored.

The second paradigm fine-tunes pretrained LLMs with supervised training. Representative works employ unified instruction templates (Lu et al., 2022), incorporate Chain-of-Thought (Huang et al., 2026), or decompose NER into subtasks (Guo et al., 2025b; Lu et al., 2024). For domain adaptation, methods leverage data augmentation (Huang et al., 2025a; Bogdanov et al., 2024), inject guiding knowledge (Yang et al., 2025; Sainz et al., 2024), or denoise pseudo-labels (Ding et al., 2024).

Retrieval-Augmented Generation. RAG integrates external information to supplement knowledge and reduce hallucination (Zhao et al., 2026; Peng et al., 2025; Ram et al., 2023). Recent advances optimize retrieval methods (Yan et al., 2024; Gao et al., 2023a; Jeong et al., 2024), perform document refinement (Trivedi et al., 2023; Liu et al., 2026), or design multi-round mechanisms (Shao et al., 2023). However, most RAG systems rely on static retrieval strategies that may not adapt to diverse query types.

Reinforcement Learning for LLMs. Reinforcement learning has been applied to LLMs (Ouyang et al., 2022), though early methods face tremendous training challenges (Schulman et al., 2017). Alternative approaches maximize reward by solving a classification problem (Rafailov et al., 2023) or optimize through advantage normalization (Hu, 2025). RL has been extended to machine translation (He

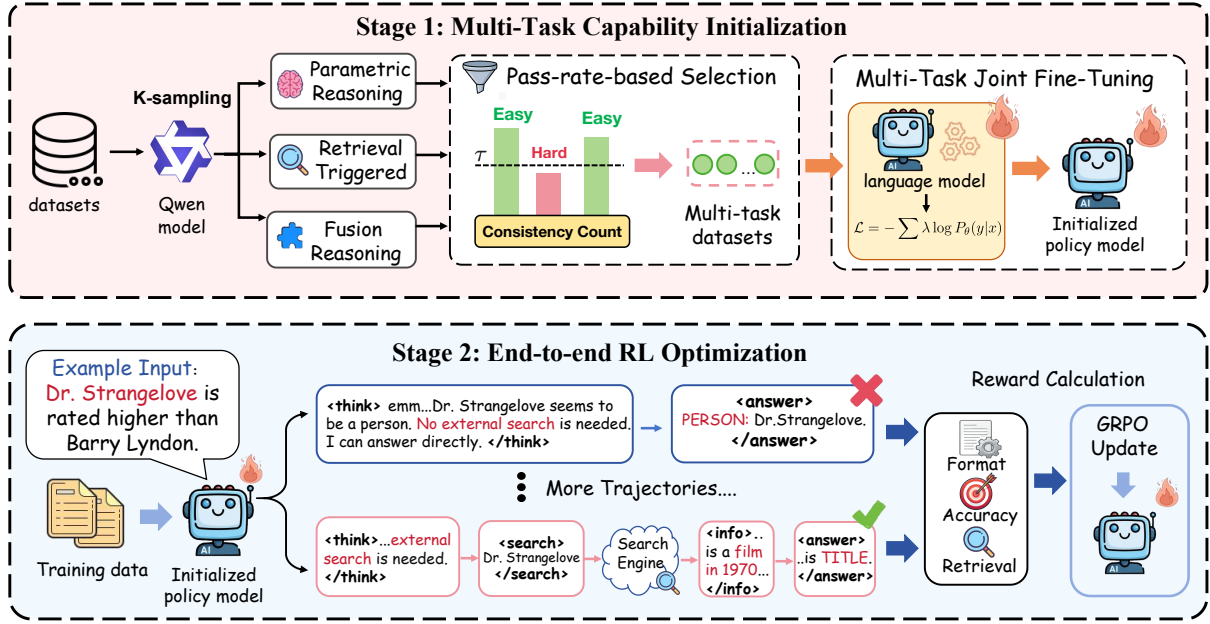


Figure 2: Overview of the NE-R1 framework. The pipeline consists of two main stages: (1) **Multi-Task Capability Initialization**, where we construct three instruction tuning tasks via pass-rate-based data selection to equip the model with foundational capabilities for parametric inference, retrieval triggering, and fusion inference; and (2) **End-to-End RL Optimization**, which employs Group Relative Policy Optimization (GRPO) to fine-tune the model, leveraging Chain-of-Thought (CoT) guided reasoning and a multi-dimensional reward mechanism to achieve adaptive retrieval and precise entity extraction.

et al., 2025; Feng et al., 2025), text-to-SQL (Pourreza et al., 2025), and question-answering (Song et al., 2025; Zheng et al., 2025).

3 Methodology

In this section, we present the NE-R1 framework. As shown in Figure 2, it comprises two key components: (1) **Multi-Task Capability Initialization (MTCI)**, which equips the model with capabilities in parametric inference, retrieval triggering, and evidence-fusion inference via multi-task instruction tuning (Section 3.1); and (2) **End-to-End RL Optimization**, which designs a Chain-of-Thought (CoT)-guided reasoning paradigm and a multi-dimensional reward, and employs Group Relative Policy Optimization (GRPO) for end-to-end optimization (Section 3.2).

3.1 Multi-Task Capability Initialization

We aim to enable adaptive retrieval decisions for NER. However, uninitialized models cannot generate effective queries or integrate retrieved results, making direct reinforcement learning inefficient and unstable. We therefore apply multi-task instruction tuning to inject three foundational capabilities into Qwen2.5-7B, using supervised data distilled

from a stronger teacher model, Qwen2.5-32B.

Task Design and Data Construction. We design three complementary instruction tuning tasks: parametric inference (T_{param}), retrieval triggering (T_{trigger}), and fusion inference (T_{rag}).

To construct high-quality training data that balances signal quality with sufficient exploration potential, we employ a pass-rate-based data selection mechanism. For each training sample, we use Qwen2.5-32B to generate $K=10$ candidate responses and apply task-specific selection criteria.

• **Parametric Inference** (T_{param}) trains the model to extract entities using only the input context and its internal parametric knowledge. The target sequence contains an <answer> block with the predicted entities. For each training sample x , we estimate the non-retrieval pass rate as

$$c_{\text{param}}(x) = \frac{1}{K} \sum_{k=1}^K \mathbb{I}(\hat{y}_{\text{param},k} = y^*), \quad (1)$$

where $\hat{y}_{\text{param},k}$ is the k -th sampled output, y^* is the gold entity annotation, and $\mathbb{I}(\cdot)$ is the indicator function. Samples with reliable parametric predictions are retained:

$$\mathcal{D}_{\text{param}} = \{x \mid c_{\text{param}}(x) \geq \tau_{\text{param}}\}. \quad (2)$$

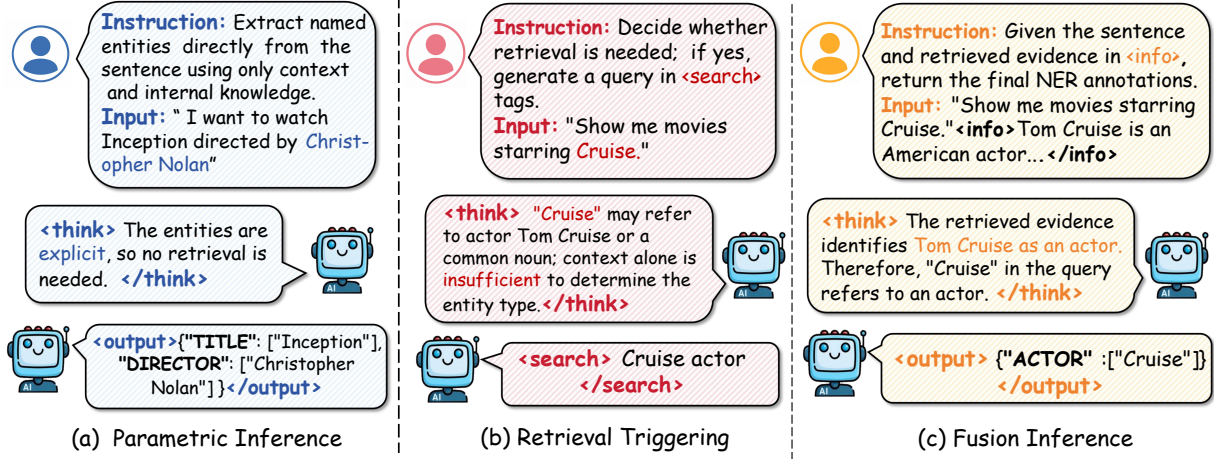


Figure 3: Illustration of the three supervised tasks in Multi-Task Capability Initialization (MTCI). (a) **Parametric Inference** performs NER directly from the input context and the model’s parametric knowledge. (b) **Retrieval Triggering** determines whether external retrieval is necessary and, if needed, generates a query in `<search>` tags. (c) **Fusion Inference** integrates retrieved evidence to produce the final NER output while preserving spans from the original input.

- **Retrieval Triggering** (T_{trigger}) trains the model to decide whether external knowledge is needed and, when necessary, to generate a rewritten search query. The supervised target contains a `<think>` fragment followed by a `<search>` block. The `<think>` fragment describes the source of uncertainty (e.g., entity ambiguity, rare mentions, or insufficient contextual evidence), while the `<search>` block contains an entity-centric rewritten query rather than the original sentence. Since no gold query exists for this task, we evaluate trigger quality indirectly via the downstream fusion pass rate (defined below) and retain trajectories whose retrieved evidence enables reliable fusion.

- **Fusion Inference** (T_{rag}) trains the model to produce the final NER prediction after retrieved evidence is inserted via `<information>` tags. Given a query generated by T_{trigger} , the retriever returns evidence, and the teacher samples K fusion outputs per training sample. We compute the retrieval-augmented pass rate as

$$c_{\text{rag}}(x) = \frac{1}{K} \sum_{k=1}^K \mathbb{I}(\hat{y}_{\text{rag},k} = y^*), \quad (3)$$

where $\hat{y}_{\text{rag},k}$ is the k -th sampled fusion output. The fusion dataset is selected by

$$\mathcal{D}_{\text{rag}} = \{x \mid c_{\text{rag}}(x) \geq \tau_{\text{rag}}\}. \quad (4)$$

Here, τ_{param} , τ_{rag} , and τ_{trigger} are task-specific selection thresholds. Beyond data filtering, the pass rates c_{param} and c_{rag} also partition training samples

into easy and hard groups, which are used to define the retrieval-benefit reward in Section 3.2.

Multi-Task Joint Fine-Tuning. As illustrated in Figure 3, we perform multi-task joint fine-tuning on the filtered samples to acquire the three capabilities. We consolidate the tasks into $\mathcal{T} = \{T_{\text{param}}, T_{\text{trigger}}, T_{\text{rag}}\}$ with distinct supervision: T_{param} generates answers using internal knowledge; T_{trigger} generates retrieval queries; and T_{rag} performs inference conditioned on retrieved information. The task-specific prompts are provided in Appendix A.4.

We jointly optimize over this mixed distribution by minimizing the weighted negative log-likelihood loss:

$$\mathcal{L} = - \sum_{(x,y) \in \mathcal{T}} \lambda_{\text{type}} \sum_{t=1}^{|y|} \log P_{\theta}(y_t \mid y_{<t}, x), \quad (5)$$

where λ_{type} denotes the task-specific balancing weight, x is the input context, y is the corresponding target output sequence, and $|y|$ is its length.

3.2 End-to-end RL Optimization

MTCI equips the model with the abilities to reason, search, and fuse evidence, but it does not directly optimize the global trade-off between NER accuracy and retrieval cost. We therefore further optimize the initialized policy with rule-based rewards under a unified trajectory-level objective.

CoT-Guided Adaptive Retrieval. To enable dynamic selection between parametric and retrieval-augmented reasoning, we design a Chain-of-Thought (CoT) guided paradigm that formalizes NER as a "reasoning-then-action" branching process. Given input x , the generation process is:

$$y = \begin{cases} \pi_\theta(y | x, c), & a(c) = \text{direct} \\ \pi_\theta(y | x, c, \mathcal{R}(q)), & a(c) = \text{search} \end{cases} \quad (6)$$

where c denotes the CoT reasoning fragment, $a(c)$ is the decision action, q is the retrieval query, and $\mathcal{R}(q)$ denotes retrieved knowledge. This comprises two steps:

- *Introspective Planning.* The model generates a <think> fragment c to analyze entity ambiguity, domain knowledge needs, and context sufficiency. With insufficient internal knowledge, it generates a <search> tag with a rewrite search query q ; otherwise, it proceeds directly to answer generation.
- *Evidence Fusion.* When <search> is triggered, search engine $\mathcal{R}$ provides top- k documents $\mathcal{R}(q)$. The model fuses input x with $\mathcal{R}(q)$ then generates entity predictions y in the <answer> block.

Multi-Dimensional Reward. We define a reward function that jointly encourages valid trajectories, accurate entity extraction, and selective retrieval.

- **Format Reward** r_{fmt} ensures that generated trajectories are parsable and verifiable:

$$r_{\text{fmt}}(y) = (\omega_c + \omega_e) \mathbb{I}[\mathcal{F}(y)] - \omega_e, \quad (7)$$

where $\mathcal{F}(y)$ checks whether tags such as <search> and <answer> are complete and properly paired, and $\omega_c, \omega_e > 0$ are the reward and penalty weights, respectively.

- **Accuracy Reward** r_{acc} uses the entity-level Micro-F1 between the prediction y and the gold annotation y^* :

$$r_{\text{acc}}(y, y^*) = \text{Micro-F1}(y, y^*). \quad (8)$$

- **Retrieval Benefit Reward** r_{bnf} encourages retrieval on hard instances while discouraging unnecessary or harmful retrieval on easy instances:

$$r_{\text{bnf}}(x, y) = \begin{cases} +w_{\text{hard}}^{\text{corr}}, & r_{\text{acc}} > \tau \wedge x \in \mathcal{X}_{\text{hard}}, \\ +w_{\text{easy}}^{\text{corr}}, & r_{\text{acc}} > \tau \wedge x \in \mathcal{X}_{\text{easy}}, \\ -w_{\text{hard}}^{\text{err}}, & r_{\text{acc}} \leq \tau \wedge x \in \mathcal{X}_{\text{hard}}, \\ -w_{\text{easy}}^{\text{err}}, & r_{\text{acc}} \leq \tau \wedge x \in \mathcal{X}_{\text{easy}}. \end{cases} \quad (9)$$

We set $w_{\text{hard}}^{\text{corr}} > w_{\text{easy}}^{\text{corr}}$ and $w_{\text{hard}}^{\text{err}} < w_{\text{easy}}^{\text{err}}$, so that successful retrieval is rewarded more on hard instances while erroneous retrieval is penalized more on easy instances. The subsets $\mathcal{X}_{\text{hard}}$ and $\mathcal{X}_{\text{easy}}$ are determined from the MTCI pass rates. The final reward is

$$R = \alpha r_{\text{acc}} + \gamma r_{\text{bnf}} + \lambda r_{\text{fmt}}, \quad (10)$$

where α, γ , and λ control the weights of accuracy, retrieval benefit, and format validity, respectively.

RL Algorithm. We optimize the policy using Group Relative Policy Optimization (GRPO). Unlike actor-critic methods, GRPO estimates advantages from a group of sampled trajectories and requires no separate critic network. Since retrieval evidence is externally injected rather than generated by the policy, we follow R1-Searcher++ (Song et al., 2025) and mask environmental observation tokens when computing policy gradients:

$$\mathbb{M}_t = \begin{cases} 1, & y_t \in \mathcal{A}_{\text{gen}}, \\ 0, & y_t \in \mathcal{O}_{\text{env}}, \end{cases} \quad (11)$$

where $\mathcal{A}_{\text{gen}}$ denotes model-generated tokens and $\mathcal{O}_{\text{env}}$ denotes externally retrieved evidence. The masked GRPO objective is

$$J(\theta) = \mathbb{E} \left[\frac{1}{G} \sum_{i=1}^G \mathcal{L}_i(\theta) - \beta D_{\text{KL}} \right], \quad (12)$$

$$\mathcal{L}_i(\theta) = \frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \mathbb{M}_t \min(\rho_{i,t} \hat{A}_i, \bar{\rho}_{i,t} \hat{A}_i), \quad (13)$$

where $\bar{\rho}_{i,t} = \text{clip}(\rho_{i,t}, 1 - \epsilon, 1 + \epsilon)$, G is the number of sampled trajectories, $\rho_{i,t}$ is the importance sampling ratio, $\hat{A}_i$ is the group-relative advantage, and β is the KL penalty coefficient.

4 Experiments

In this section, we systematically evaluate the proposed NE-R1 framework to address the following research questions:

- **RQ1:** Does NE-R1 outperform existing models on NER benchmarks?
- **RQ2:** Does NE-R1 generalize cross-domain in zero-shot?
- **RQ3:** What is the contribution of each NE-R1 component?
- **RQ4:** Does NE-R1 learn effective adaptive retrieval behavior?

Category	Model	MIT-Movie	MIT-Rest.	OntoNotes	GENIA	Avg.
Few-shot LLMs	GPT-4o	61.49	59.90	64.29	44.63	57.58
	Llama-3.3-70B	60.76	53.62	59.98	41.91	54.07
	Qwen3-32B	52.14	56.47	46.71	39.99	48.83
	DeepSeek-R1-MoE	63.93	64.87	67.81	53.81	62.61
	GPT-5	61.23	66.53	72.28	51.22	62.81
NER baselines	InstructUIE (Wang et al., 2023)	89.58	82.59	88.64	75.71	84.13
	GLiNER-L (Zaratiana et al., 2024)	87.90	83.60	89.80	<u>78.40</u>	84.93
	B^2 NER (Yang et al., 2025)	<u>90.78</u>	<u>83.71</u>	84.31	76.43	83.81
	ReasoningNER (Huang et al., 2026)	89.60	81.40	89.10	78.30	84.60
	UniNER-7B (Zhou et al., 2024)	90.17	82.35	<u>89.91</u>	77.54	<u>85.00</u>
Retrieval	Standard SFT (No Search)	85.24	80.68	83.47	75.92	80.83
	RAG (Always Search)	85.56	81.72	84.38	77.23	82.22
NE-R1	NE-R1 (MTCI only)	87.94	83.46	85.12	77.77	83.57
	NE-R1 (full)	91.05 ^{+0.27}	85.44 ^{+1.73}	91.89 ^{+1.98}	81.70 ^{+3.30}	87.52 ^{+2.52}

Table 1: In-domain NER results. We report Micro-F1. Bold marks the best result in each column, underlining marks the strongest prior baseline, and the highlighted NE-R1 row additionally reports gains over the strongest prior result.

4.1 Setup

Datasets and Metrics. We evaluate our model on standard benchmarks for both in-domain and cross-domain scenarios. For in-domain evaluation, we use OntoNotes 5.0 (Weischedel et al., 2013), MIT-Movie (Liu et al., 2013), MIT-Restaurant (Liu et al., 2013), and GENIA (Kim et al., 2003), covering general, short-text, and specialized domains. For cross-domain evaluation (RQ2), following the standard protocol (Nandi and Agrawal, 2024), models are trained on CoNLL-03 and evaluated zero-shot on the five CrossNER domains. All results are evaluated using strict entity-level Micro-F1. Detailed dataset statistics are provided in Appendix A.1.

Baselines. We benchmark NE-R1 against general-purpose LLMs (GPT-4o, GPT-5, Llama-3-70B, Qwen3-32B, DeepSeek-R1) and domain-specific NER models (InstructUIE, UniNER, GLiNER, B^2 NER, ReasoningNER) for in-domain evaluation. For cross-domain evaluation, we follow (Nandi and Agrawal, 2024) and adopt their baselines including transfer-based, prompting-based, and retrieval-augmented approaches.

Implementation Details. We use Qwen-2.5-7B as the backbone pre-trained language model, Qwen-2.5-32B generates training contents for MTCI. All experiments are conducted on 8×NVIDIA A800 GPUs. In MTCI phase, we set learning rate to 1e-5

and batch size to 16. In RL phase, we set the rollout size to 5 and the learning rate to 1e-6. We employ a low-variance KL penalty (coefficient 1e-3) and a clip coefficient of 0.2. We use multi-source web search engines as external knowledge acquisition channels. For more detailed implementation details, refer to the appendix A.2.

4.2 Main Results (In-Domain Evaluation)

We first evaluate NE-R1 under the standard supervised setting. We compare the model against comparable (7B) and larger (70B+) baselines across four datasets characterized by diverse domain distributions and knowledge dependencies. As shown in Table 1, NE-R1 achieves superior performance on all benchmarks, consistently outperforming both equal-sized and larger models. Specifically, On MIT-Movie and MIT-Restaurant, it leads among 7B models with **91.05** and **85.44** F1 score, on the general-domain OntoNotes 5.0, it attains **91.89** F1 score, surpassing InstructUIE and UniNER-7B by **over 2.0 points**. On the terminology-dense GENIA, it achieves **81.70** F1 score, a significant **3 point improvement** over GLiNER. These results demonstrate the effectiveness of adaptive retrieval across the four evaluated in-domain NER benchmarks.

Model	Politics	Natural Science	Music	Literature	AI	Avg.
BiLSTM-CRF (Lample et al., 2016)	56.60	49.97	44.79	43.03	43.56	47.59
Coach (Liu et al., 2020)	61.50	52.09	51.66	48.35	45.15	51.75
CROSS-DOMAIN LM (Jia et al., 2019)	68.44	64.31	63.56	59.59	53.70	61.92
FLAIR (Akbik et al., 2019)	69.54	64.71	65.60	61.35	52.48	62.73
BARTNER (Liu et al., 2021)	69.90	65.14	65.35	58.93	53.00	62.46
LIGHTNER (Chen et al., 2022)	72.78	66.74	72.28	65.17	35.82	62.56
LST-NER (Zheng et al., 2022)	73.25	70.07	76.83	70.76	63.28	70.84
LANER (Hu et al., 2022)	74.06	71.83	78.78	71.11	65.79	72.31
CP-NER (Chen et al., 2023)	74.25	75.82	79.10	72.17	67.95	73.86
GPT-NER (Wang et al., 2025a)	74.71	70.77	78.30	62.18	66.07	70.41
PromptNER (GPT-3.5) (Ashok and Lipton, 2023)	71.74	64.83	77.78	64.15	59.35	67.57
PromptNER (GPT-4) (Ashok and Lipton, 2023)	78.61	72.59	84.26	74.44	64.83	74.95
RAG + GPT-4 (sentence embeddings)	78.20	73.52	83.61	71.32	66.91	74.71
RAG + GPT-4 (word embeddings)	78.63	73.95	84.25	74.68	68.19	75.94
IF-WRANER (Nandi and Agrawal, 2024)	<u>79.80</u>	<u>75.31</u>	85.43	<u>75.52</u>	<u>68.81</u>	<u>76.97</u>
NE-R1 (ours)	81.35 ^{+1.55}	75.83 ^{+0.52}	83.51 ^{-1.92}	76.30 ^{+0.78}	73.78 ^{+4.97}	78.15 ^{+1.18}

Table 2: Zero-shot cross-domain NER results on CrossNER. We report Micro-F1. Bold marks the best result in each column, underlining marks the strongest prior baseline, and the highlighted NE-R1 row reports gains over IF-WRANER, the strongest overall baseline in average F1.

Variant	Movie	Res.	Onto.	Gen.	Avg.
Base model	51.50	56.53	44.24	39.96	48.06
w/o RL	87.94	83.46	85.12	77.77	83.57
w/o MTCI	71.92	67.84	74.63	67.58	70.49
w/o $T_{\text{trigger}} \& T_{\text{rag}}$	84.77	80.06	82.76	75.34	80.73
w/o CoT	89.84	84.73	89.37	80.71	86.16
w/o r_{bnf}	88.71	83.68	86.64	79.39	84.61
NE-R1 (full)	91.05	85.44	91.89	81.70	87.52

Table 3: Ablation study on different datasets. All results are micro-F1 score. Movie, Res., Onto., and Gen. denote MIT Movie, MIT Restaurant, OntoNotes, and GENIA datasets, respectively. “w/o” denotes removing a specific component from our full NE-R1 model.

4.3 Cross-Domain Evaluation

To address RQ2, we train NE-R1 on CoNLL-03 and evaluate it zero-shot on the five CrossNER domains (Table 2). Compared with IF-WRANER, NE-R1 improves Politics, Natural Science, Literature, and AI by 1.55, 0.52, 0.78, and 4.97 F1 points, respectively, but performs 1.92 points lower on Music. It nevertheless achieves the best five-domain average of **78.15**, outperforming IF-WRANER by **1.18** F1 points.

The largest gain on AI suggests that retrieval is particularly useful for specialized and evolving technical entities. In contrast, the smaller gains on Natural Science and Literature and the decrease

on Music indicate that retrieval is less effective for errors involving local context, entity boundaries, aliases, or dataset-specific label distinctions, and may introduce noise for ambiguous entity names. Overall, the benefit of adaptive retrieval varies across domains.

4.4 Ablation Studies

To dissect the efficacy of NE-R1 (RQ3), we performed ablations on training stages, initialization tasks, and core components (Table 3).

Training Stage Ablation. *w/o MTCI* removes the Multi-Task Capability Initialization stage; *w/o end-to-end RL* removes the RL Optimization stage. Removing MTCI prevents effective query generation and knowledge integration, while removing end-to-end RL causes notable performance drops, confirming both stages are essential.

Initialization Task Ablation. *w/o T_{trigger} and T_{rag}* excludes Retrieval Triggering and Fusion Inference from initialization tasks. This causes a 6.8 point drop in F1 score (87.52→80.73), demonstrating that joint training on query generation and information fusion is crucial.

Reasoning and Reward Ablation. *w/o CoT* removes Chain-of-Thought reasoning; *w/o r_{bnf}* removes Retrieval Benefit Reward. Ablating r_{bnf} proves more detrimental than removing CoT, indi-

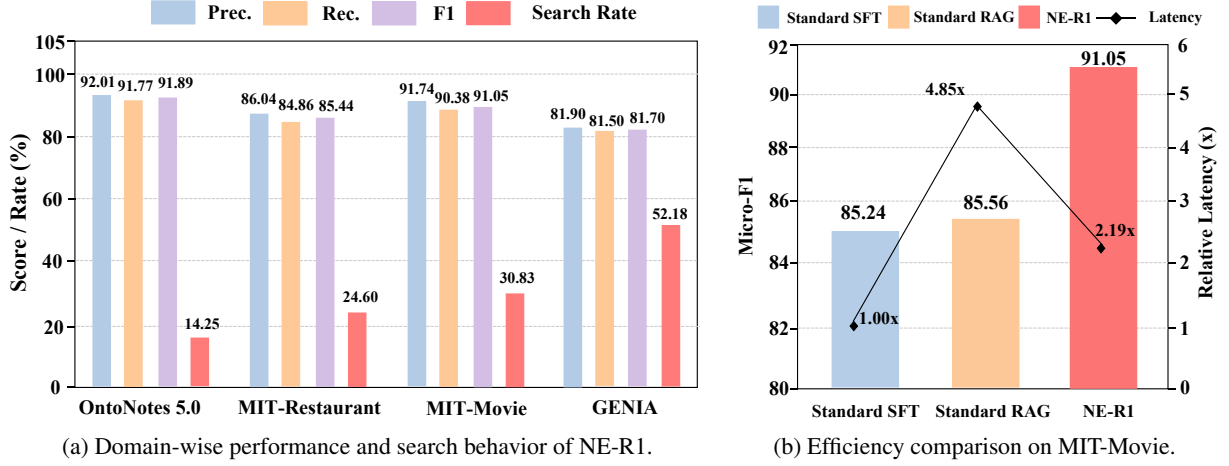


Figure 4: Analysis of NE-R1’s adaptive retrieval behavior. (a) NE-R1 calibrates retrieval frequency based on domain knowledge density. (b) NE-R1 achieves the best trade-off between accuracy and inference cost on MIT-Movie.

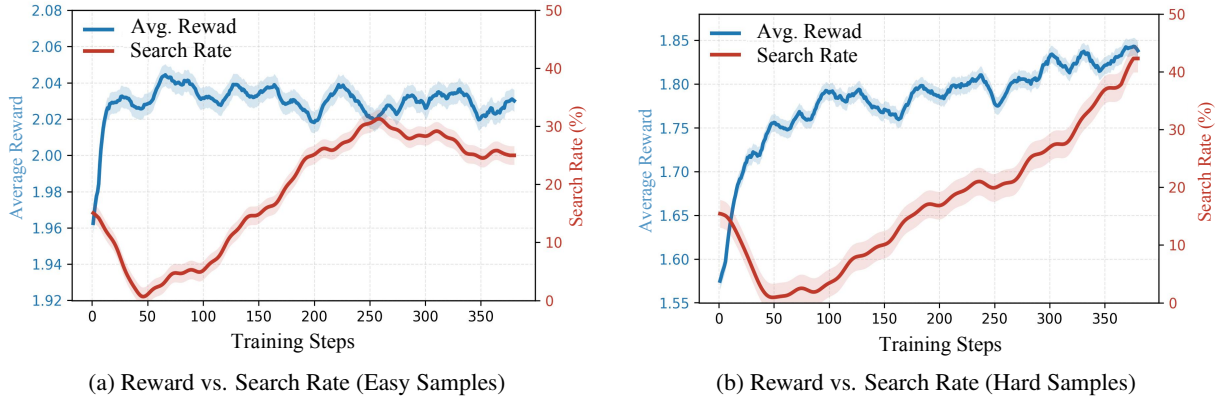


Figure 5: Reward and search rate dynamics under different sample difficulties. Note that the blue line represents Average Reward and the red line represents Search Rate.

cating the benefit-aware reward is central to balancing retrieval costs.

4.5 Analysis of Adaptive Search Behavior

To address RQ4, we analyze the retrieval behavior of NE-R1 from three perspectives: domain-level retrieval frequency, the efficiency–performance trade-off, and training dynamics. Together, these analyses examine whether retrieval decisions vary with knowledge demand rather than following a fixed always-retrieve or never-retrieve policy.

Search Behavior Divergence. Figure 4(a) compares retrieval rates across domains with different knowledge requirements. The retrieval rate is 14.25% on OntoNotes 5.0, where many entities can be recognized from local context and parametric knowledge. It increases to 24.60% on MIT-Restaurant and 30.83% on MIT-Movie, whose short queries often contain ambiguous names with limited contextual information. On GENIA, the

rate reaches 52.18%, reflecting the greater need for external knowledge when recognizing specialized biomedical terminology. This variation suggests that NE-R1 adjusts its retrieval frequency to the knowledge demands of different domains instead of applying a uniform retrieval policy.

Speed–Quality Trade-off. We compare NE-R1 with Standard SFT and Standard RAG (always-retrieve) on MIT-Movie, as shown in Figure 4(b). Standard SFT achieves 85.24 Micro-F1 with a relative latency of 1.00 $\times$, whereas Standard RAG achieves only a modest improvement to 85.56 Micro-F1 while increasing latency to 4.85 $\times$. In comparison, NE-R1 achieves the highest Micro-F1 score of 91.05 with a relative latency of 2.19 $\times$, corresponding to an approximately 55% reduction in inference latency relative to Standard RAG. These results show that adaptive retrieval avoids part of the overhead of indiscriminate retrieval while obtaining more useful external evidence.

Reward Evolution. Figure 5 illustrates the evolution of average reward and retrieval rate for easy and hard samples. For easy samples (Figure 5a), the retrieval rate decreases while the reward remains high, indicating that the model increasingly relies on parametric knowledge and avoids unnecessary retrieval. For hard samples (Figure 5b), the retrieval rate increases as training progresses, allocating more retrieval to inputs for which external evidence is more likely to be useful. The opposing trends show that training does not simply encourage a globally higher or lower retrieval rate. Instead, they support that NE-R1 learns a difficulty-aware retrieval strategy conditioned on the expected utility of external evidence.

4.6 Retrieval Leakage Audit

NE-R1 does not directly retrieve test answers or benchmark annotations. The model generates entity-centric queries around uncertain mentions only when external knowledge is needed, rather than submitting complete test sentences to search engines. The retrieved content provides entity-related background knowledge for semantic disambiguation, while predicted entity spans and types remain constrained by the original input and the predefined label set.

We conduct an exhaustive leakage audit across all four in-domain benchmarks: MIT-Movie, MIT-Restaurant, OntoNotes 5.0, and GENIA. Their test sets contain 1,953, 1,521, 8,262, and 1,854 instances, respectively, covering 13,590 test instances in total. For every instance that triggers retrieval, we manually inspect the generated query and all of its top-3 retrieved snippets. The exhaustive audit covers **3,120** retrieval-triggered instances and **9,360** query-snippet pairs in total.

We examine whether the retrieved content contains annotation-related terms, BIO tags or other structured NER outputs, gold entity lists, complete test sentences, or links to benchmark pages and annotation files. We find no cases in which retrieval causes test-answer leakage.

5 Conclusion

In this paper, we introduced NE-R1, an adaptive retrieval-augmented framework for NER that implements a “**retrieval-on-demand**” mechanism. NE-R1 first acquires parametric inference, retrieval triggering, and evidence-fusion capabilities through multi-task instruction tuning, and then

jointly optimizes NER predictions and retrieval decisions through end-to-end reinforcement learning. This design allows the model to rely on parametric knowledge for familiar inputs and retrieve external evidence for knowledge-intensive or ambiguous entities.

Across the four evaluated in-domain benchmarks, NE-R1 improves the average F1 score by 2.52 points over the strongest overall baseline. In zero-shot transfer from CoNLL-03 to CrossNER, it improves performance in four of the five domains and achieves the best five-domain average, exceeding the strongest baseline by 1.18 F1 points. Behavioral analysis further shows that retrieval varies across domains and sample difficulties, while on MIT-Movie, NE-R1 reduces inference latency by approximately 55% relative to always-retrieve RAG. Overall, these results demonstrate that NE-R1 consistently outperforms existing generative and retrieval-augmented NER methods with higher efficiency, while generalizing robustly to unseen entity types.

Limitations

First, NE-R1 relies on live, multi-source Web search. The retrieved content and its ranking may change over time because of API updates, indexing changes, and regional variation. Consequently, exact reproduction with live search services cannot be guaranteed, even under the same retrieval configuration.

Second, our empirical conclusions are limited to the evaluated English NER benchmarks and the current multi-source Web-search setting. The experiments do not establish that NE-R1 generalizes to other languages, domains, or retrieval backends. Moreover, heterogeneous search engines may return multilingual, conflicting, or uneven-quality evidence, and we do not separately quantify the effects of these factors.

Finally, the current framework supports only single-turn retrieval. While iterative retrieval may provide additional evidence for obscure or highly ambiguous entities, it may also introduce additional latency, noise, and error propagation. The main results follow the standard strict entity-level Micro-F1 evaluation protocol. Because repeated-run results or test predictions are unavailable for many comparison systems, we cannot conduct a uniform run-level variance or paired significance analysis across all baselines.

References

- Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)*, pages 54–59.
- Dhananjay Ashok and Zachary C Lipton. 2023. Promptner: Prompting for named entity recognition. *arXiv preprint arXiv:2305.15444*.
- Sergei Bogdanov, Alexandre Constantin, Timothée Bernard, Benoit Crabbé, and Etienne P Bernard. 2024. Nuner: Entity recognition encoder pre-training via llm-annotated data. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11829–11841.
- Xiang Chen, Lei Li, Shumin Deng, Chuanqi Tan, Changliang Xu, Fei Huang, Luo Si, Huajun Chen, and Ningyu Zhang. 2022. Lightner: A lightweight tuning paradigm for low-resource ner via pluggable prompting. In *Proceedings of the 29th international conference on computational linguistics*, pages 2374–2387.
- Xiang Chen, Lei Li, Shuofei Qiao, Ningyu Zhang, Chuanqi Tan, Yong Jiang, Fei Huang, and Huajun Chen. 2023. One model for all domains: collaborative domain-prefix tuning for cross-domain ner. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 5030–5038.
- Zihan Chen, Lei Shi, Weize Wu, Qiji Zhou, and Yue Zhang. 2025. Allabel: Three-stage active learning for llm-based entity recognition using demonstration retrieval. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25173–25187.
- 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.
- Zhuojun Ding, Wei Wei, Xiaoye Qu, and Dangyang Chen. 2024. Improving pseudo labels with global-local denoising framework for cross-lingual named entity recognition. *arXiv preprint arXiv:2406.01213*.
- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128.
- Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 6491–6501.
- Zhaopeng Feng, Shaosheng Cao, Jiahao Ren, Jiayuan Su, Ruizhe Chen, Yan Zhang, Zhe Xu, Yao Hu, Jian Wu, and Zuozhu Liu. 2025. Mt-r1-zero: Advancing llm-based machine translation via r1-zero-like reinforcement learning. *arXiv preprint arXiv:2504.10160*.
- Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023a. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2023b. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.
- Hila Gonen, Srinu Iyer, Terra Blevins, Noah A Smith, and Luke Zettlemoyer. 2023. Demystifying prompts in language models via perplexity estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10136–10148.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shiron Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, et al. 2025a. [DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning](#). *Nature*, 645(8081):633–638.
- Quanjian Guo, Yihong Dong, Ling Tian, Zhao Kang, Yu Zhang, and Sijie Wang. 2025b. Baner: Boundary-aware llms for few-shot named entity recognition. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10375–10389.
- Mingui He, Yilun Liu, Shimin Tao, Yuanchang Luo, Hongyong Zeng, Chang Su, Li Zhang, Hongxia Ma, Daimeng Wei, Weibin Meng, Hao Yang, Boxing Chen, and Osamu Yoshie. 2025. [R1-t1: Fully incentivizing translation capability in llms via reasoning learning](#). *Preprint*, arXiv:2502.19735.
- Jian Hu. 2025. Reinforce++: A simple and efficient approach for aligning large language models. *arXiv e-prints*, pages arXiv–2501.
- Jinpeng Hu, He Zhao, Dan Guo, Xiang Wan, and Tsung-Hui Chang. 2022. A label-aware autoregressive framework for cross-domain ner. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2222–2232.
- Hui Huang, Yanping Chen, Ruizhang Huang, Chuan Lin, and Yongbin Qin. 2026. A reasoning paradigm for named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 31140–31148.

- Li Huang, Haowen Liu, Qiang Gao, Jiajing Yu, Guisong Liu, and Xueqin Chen. 2025a. Adversity-aware few-shot named entity recognition via augmentation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 24132–24140.
- Shizhou Huang, Bo Xu, Yang Yu, Changqun Li, and Xin Alex Lin. 2025b. Guidener: Annotation guidelines are better than examples for in-context named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 24159–24166.
- Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. 2024. [Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity](#). 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 7036–7050, Mexico City, Mexico. Association for Computational Linguistics.
- Chen Jia, Xiaobo Liang, and Yue Zhang. 2019. Cross-domain ner using cross-domain language modeling. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 2464–2474.
- Katie Kang, Eric Wallace, Claire Tomlin, Aviral Kumar, and Sergey Levine. 2025. Unfamiliar finetuning examples control how language models hallucinate. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3600–3612.
- J-D Kim, Tomoko Ohta, Yuka Tateisi, and Jun’ichi Tsujii. 2003. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180–i182.
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 260–270.
- Yanchun Li, Senlin Deng, Dongsu Shen, Shujuan Tian, and Saiqin Long. 2024. Wkner: Enhancing named entity recognition with word segmentation constraints and knn retrieval. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17651–17663.
- 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. IEEE.
- Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM computing surveys*, 55(9):1–35.
- Shuyi Liu, Yu-Ming Shang, and Xi Zhang. 2026. Truthfulrag: Resolving factual-level conflicts in retrieval-augmented generation with knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 32168–32176.
- Zihan Liu, Genta Indra Winata, Peng Xu, and Pascale Fung. 2020. Coach: A coarse-to-fine approach for cross-domain slot filling. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 19–25.
- 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. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13452–13460.
- Jinghui Lu, Ziwei Yang, Yanjie Wang, Xuejing Liu, Brian Mac Namee, and Can Huang. 2024. Padellmner: parallel decoding in large language models for named entity recognition. *Advances in neural information processing systems*, 37:117853–117880.
- Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified structure generation for universal information extraction. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 5755–5772.
- Xuezhe Ma and Eduard H. Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074.
- Subhadip Nandi and Neeraj Agrawal. 2024. Improving few-shot cross-domain named entity recognition by instruction tuning a word-embedding based retrieval augmented large language model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 686–696.
- 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.
- Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. 2025. Graph retrieval-augmented generation: A survey. *ACM Transactions on Information Systems*, 44(2):1–52.

- Mohammadreza Pourreza, Shayan Talaei, Ruoxi Sun, Xingchen Wan, Hailong Li, Azalia Mirhoseini, Amin Saberi, and Sercan Ö. Arik. 2025. [Reasoning-sql: Reinforcement learning with sql tailored partial rewards for reasoning-enhanced text-to-sql](#). *Preprint*, arXiv:2503.23157.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.
- Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.
- Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lacalle, German Rigau, and Eneko Agirre. 2024. Gollie: Annotation guidelines improve zero-shot information-extraction. In *International Conference on Learning Representations*, volume 2024, pages 47083–47107.
- Erik Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 142–147.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Wei Liang Seow, Iti Chaturvedi, Amber Hogarth, Rui Mao, and Erik Cambria. 2025. A review of named entity recognition: from learning methods to modelling paradigms and tasks. *Artificial Intelligence Review*, 58(10):315.
- Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9248–9274.
- Huatong Song, Jinhao Jiang, Wenqing Tian, Zhipeng Chen, Yuhuan Wu, Jiahao Zhao, Yingqian Min, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. 2025. R1-searcher++: Incentivizing the dynamic knowledge acquisition of llms via reinforcement learning. *arXiv preprint arXiv:2505.17005*.
- Zeqi Tan, Shen Huang, Zixia Jia, Jiong Cai, Yinghui Li, Weiming Lu, Yueting Zhuang, Kewei Tu, Pengjun Xie, Fei Huang, and Yong Jiang. 2023. [DAMO-NLP at SemEval-2023 task 2: A unified retrieval-augmented system for multilingual named entity recognition](#). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2014–2028, Toronto, Canada. Association for Computational Linguistics.
- Zeliang Tong, Zhuojun Ding, and Wei Wei. 2025. Evo-prompt: Evolving prompts for enhanced zero-shot named entity recognition with large language models. In *Proceedings of the 31st international conference on computational linguistics*, pages 5136–5153.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 10014–10037.
- Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025a. Gpt-ner: Named entity recognition via large language models. In *Findings of the association for computational linguistics: NAACL 2025*, pages 4257–4275.
- Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, Jihua Kang, Jingsheng Yang, Siyuan Li, and Chunsai Du. 2023. [Instructuie: Multi-task instruction tuning for unified information extraction](#). *Preprint*, arXiv:2304.08085.
- Zhenhua Wang, Huiru Chen, Guang Xu, and Ming Ren. 2025b. A novel large-language-model-driven framework for named entity recognition. *Information Processing & Management*, 62(3):104054.
- Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. [OntoNotes release 5.0 LDC2013T19](#). Web Download. Philadelphia: Linguistic Data Consortium. LDC Catalog No.: LDC2013T19.
- Tingyu Xie, Qi Li, Jian Zhang, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2023. Empirical study of zero-shot ner with chatgpt. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 7935–7956.
- Tingyu Xie, Qi Li, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2024. Self-improving for zero-shot named entity recognition with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 583–593.
- Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*.
- Yuming Yang, Wantong Zhao, Caishuang Huang, Junjie Ye, Xiao Wang, Huiyuan Zheng, Yang Nan, Yuran Wang, Xueying Xu, Kaixin Huang, Yunke Zhang, Tao Gui, Qi Zhang, and Xuanjing Huang. 2025. [Beyond boundaries: Learning a universal entity taxonomy across datasets and languages for open named](#)

entity recognition. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10902–10923, Abu Dhabi, UAE. Association for Computational Linguistics.

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.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui. 2026. Retrieval-augmented generation for ai-generated content: A survey. *Data Science and Engineering*, pages 1–29.

Junhao Zheng, Haibin Chen, and Qianli Ma. 2022. Cross-domain named entity recognition via graph matching. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2670–2680.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. 2025. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 414–431.

DAI Zhenwei, Chen Luo, Zhen Li, Xianfeng Tang, Hanqing Lu, Rahul Goutam, and Haiyang Zhang. 2024. Ra-ner: Retrieval augmented ner for knowledge intensive named entity recognition. In *The Second Tiny Papers Track at ICLR 2024*.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2024. Universalner: Targeted distillation from large language models for open named entity recognition. In *International conference on learning representations*, volume 2024, pages 12276–12294.

A Appendix

A.1 Dataset Statistics

Table 4 presents the detailed statistics of the datasets used in our experiments, including the number of sentences in train/dev/test splits, the number of entity types, the average token length, and the average number of entities per sentence.

In-Domain Datasets. We utilize a diverse set of benchmarks to assess generalizability across different data distributions:

- **Standard & General Domain: CoNLL-03** (Sang and De Meulder, 2003) and **Ontonotes 5.0** (Weischedel et al., 2013) represent standard news and general text domains. Ontonotes 5.0 is the largest dataset in our evaluation, containing over 59k training samples with 18 fine-grained entity types.
- **Specialized Domain: GENIA** (Kim et al., 2003) focuses on the biomedical domain. As shown in Table 4, it exhibits a high density of entities (Avg. entities = 3.5), reflecting the knowledge-intensive nature of scientific texts.
- **Short-Text Domain: MIT-Movie and MIT-Restaurant** (Liu et al., 2013) consist of informal user queries. These datasets have significantly shorter sentence lengths (Avg. tokens $\approx$ 9–11) compared to standard news corpora, posing challenges for context-dependent entity recognition.

Cross-Domain Datasets. For the cross-domain evaluation scenarios (trained on CoNLL-03), we employ the **CrossNER** benchmark (Liu et al., 2021). This dataset encompasses five distinct domains: *AI*, *Literature*, *Music*, *Politics*, and *Science*. While our main experiment primarily follows a zero-shot setting, Table 4 lists the standard data splits provided by the benchmark, which are typically used for low-resource or few-shot adaptation studies.

A.2 Implementation Details

Hyperparameters. In Table 5, we present detailed values of hyperparameters used in our study, categorized into the Multi-Task Capability Initialization (MTCI) phase and the Reinforcement Learning (RL) phase. All experiments were conducted on 8 NVIDIA A800 GPUs using Qwen-2.5-7B as the backbone model.

In the SFT phase, we utilized the AdamW optimizer with a learning rate of 1e-5 and a cosine learning rate scheduler. The training was conducted for 5 epochs with a warmup ratio of 0.1. We set the per-device batch size to 16 with gradient accumulation steps set to 1, and enabled bfloat16 precision to optimize memory usage.

In the RL phase, the model was further optimized using Group Relative Policy Optimization (GRPO), without introducing a critic network. We lowered the learning rate to 1e-6 with a constant scheduler and a 1% warmup phase. The training spanned 500 steps with a data collection batch size of 512, and GRPO policy updates were performed on mini-batches of size 256. The clipping ratio was set to 0.2, and we incorporated a KL divergence penalty with a coefficient of 0.001, utilizing a low-variance estimator (low_var_kl). Regarding model configuration, we limited the maximum prompt length to 4096 tokens and the response length to 500 tokens, supporting up to 2 interaction turns. During the rollout phase, we set the sampling temperature to 0.95 and generated 5 candidate responses per prompt (N=5). For retrieval-augmented generation, the top-3 relevant documents were retrieved. Some implementation frameworks retain generic policy-optimization parameter names such as mini-batch or clip-ratio settings; in our experiments, the actual advantage estimator is GRPO.

Search Engines. In the implementation of the retrieval module, we utilize multi-source web search engines (as shown in the table 6) as the external knowledge acquisition channels, and dynamically incorporate the retrieval results into the reasoning process of the model through a unified search interface.

A.3 Supplementary Experiments and Analyses

Matched-rate random trigger baseline. To further isolate the effect of the learned retrieval policy, we compare NE-R1 with random retrieval policies under different fixed search rates. As shown in Table 7, random triggering improves over naive retrieval decisions in some settings, but remains consistently inferior to the adaptive policy of NE-R1. This confirms that NE-R1 does not merely benefit from using a certain retrieval budget; instead, its learned trigger policy better identifies when external evidence is needed.

Dataset	# train	# dev	# test	# types	Avg. tokens	Avg. entities
Ontonotes(Weischedel et al., 2013)	59924	8528	8262	18	18	0.9
GENIA(Kim et al., 2003)	15023	1669	1854	5	43	3.5
MIT-Movie(Liu et al., 2013)	6816	1000	1953	12	11	2.0
MIT-Restaurant(Liu et al., 2013)	7660	1000	1521	8	9	1.7
CoNLL-03(Sang and De Meulder, 2003)	14041	3250	3453	3	25	1.9
CrossNER AI(Liu et al., 2021)	100	350	431	13	52	5.3
CrossNER Literature(Liu et al., 2021)	100	400	416	11	54	5.4
CrossNER Music(Liu et al., 2021)	100	380	465	12	57	6.5
CrossNER Politics(Liu et al., 2021)	199	540	650	8	61	6.5
CrossNER Science(Liu et al., 2021)	200	450	543	16	54	5.4

Table 4: Dataset statistics.

Hyperparameter	Value
<i>MTCI Phase</i>	
Learning Rate	1e-5
LR Scheduler	Cosine
Warmup Ratio	0.1
Num Epochs	5
Per-Device Batch Size	16
Gradient Accumulation	1
<i>RL Phase</i>	
Optimizer	AdamW
Learning Rate	1e-6
Warmup Ratio	0.01
Total Steps	500
Data Collection Batch Size	512
Clip Ratio	0.2
KL Coefficient	0.001
KL Type	low_var_kl
Rollout & Generation	
Max Prompt Length	4096
Max Response Length	500
Max Turns	2
Temperature	0.95
Rollout Samples (N)	5
Retriever Top- k	3

Table 5: Detailed hyperparameter configuration for both MTCI and RL phases.

Search Engine	URL
Baidu Search	https://www.baidu.com
Google Search	https://www.google.com
Bing Search	https://www.bing.com
Wikipedia	https://www.wikipedia.org

Table 6: Search engines used in the retrieval module. All search engines are accessed through a unified retrieval interface.

Discussion on retrieval leakage risk. A potential concern for retrieval-augmented evaluation on public NER benchmarks is that a search engine may return benchmark annotations or otherwise create an “open-book” setting. NE-R1 mitigates this risk in three ways. First, the emitted query is

not the original test sentence; it is an entity-centric rewritten query produced inside the <search> tag, such as a request for background information about an ambiguous mention. Second, the retrieved content is inserted as natural-language evidence through <information> tags and does not provide BIO labels or dataset-specific NER annotations. Third, the always-retrieve baseline does not dominate NE-R1, suggesting that merely increasing retrieval exposure is insufficient; the benefit comes from deciding when the external evidence is useful. The zero-shot cross-domain evaluation further supports that the model relies on transferable retrieval and reasoning behavior rather than memorizing benchmark-specific labels.

A.4 MIT-Movie Prompt Templates

This section presents the prompt templates used in our MTCI and evaluation pipelines, using MIT-Movie as a representative example. The same prompting scheme is applied to other datasets by adapting the task description and entity label set accordingly. Specifically, the task prompts for parametric inference, retrieval triggering, and fusion inference, together with the evaluation-time adaptive retrieval prompt, are shown in figures 6, 7, 8, and 9, respectively.

Parametric inference prompt. This prompt initializes direct NER extraction without retrieval. The examples emphasize preserving the original span text and returning a compact JSON annotation.

Retrieval triggering prompt. This prompt teaches the model to separate clear cases from ambiguous or knowledge-intensive ones, producing a search query only when external evidence is likely

Method	Search Rate	MIT-Movie	MIT-Rest.	OntoNotes	GENIA	Avg.
NE-R1 + Random	20%	88.73	84.51	88.76	78.94	85.24
NE-R1 + Random	35%	89.61	84.19	87.83	79.87	85.38
NE-R1 + Random	50%	88.12	83.44	86.71	80.63	84.73
NE-R1 (Adaptive)	14–52%	91.05	85.44	91.89	81.70	87.52

Table 7: Comparison between random retrieval triggering and NE-R1’s adaptive retrieval policy. Random triggering uses fixed retrieval rates, while NE-R1 adaptively adjusts its search rate across datasets according to input difficulty and domain knowledge requirements. All results are Micro-F1 scores.

MIT-Movie Task 1: Parametric Inference (T_{param})

You are a senior NER expert. Please extract entity spans from the given query.

Entity Categories: ACTOR: real-world actors' names; CHARACTER: movie characters; DIRECTOR: film directors' names; GENRE: movie genres, such as action, comedy, horror, mystery, and 3D; PLOT: storyline phrases, such as vampires, ghosts, war, or explosions; RATING: age-rating labels; RATINGS_AVERAGE: scores, star ratings, or recommendation levels; REVIEW: subjective review terms; SONG: song titles; TITLE: official movie titles; TRAILER: trailer or preview mentions; YEAR: years or time expressions related to movies.

Output Rules: (1) Extract entity spans according to context and keep the original surface text. (2) Return a JSON string whose keys are entity categories and whose values are span lists. (3) Omit empty categories.

Examples:
 Query: on what film did brandon lee last seen
 Output: {"ACTOR": ["brandon lee"]}
 Query: where did george clooney film the ides of march
 Output: {"ACTOR": ["george clooney"], "TITLE": ["the ides of march"]}
 Query: is there a mystery movie featuring a f silver
 Output: {"GENRE": ["mystery"], "DIRECTOR": ["a f silver"]}
 Query: list the sandlot movies
 Output: {"TITLE": ["sandlot"]}

Start: Query: what movie has the shortest trailer
 Output:

Figure 6: MIT-Movie prompt for the parametric inference task. This task trains the model to perform NER directly without external retrieval.

to help.

Fusion inference prompt. This prompt trains evidence-aware extraction. It explicitly instructs the model to use retrieved snippets for disambiguation while keeping final entity spans grounded in the original query.

Evaluation prompt. The evaluation prompt unifies the direct and retrieval-augmented paths. The model may search when necessary, but the final answer must always be a valid JSON annotation inside the answer tag.

A.5 Case Study

Successful adaptive retrieval behavior. Figure 10 visualizes the inference paths of NE-R1 across varying sample difficulties, demonstrating its capability for uncertainty-driven on-demand retrieval. In the ambiguous Case 1, the span "Crash" is polysemous (common noun vs. movie title). Recognizing that context alone is insufficient, the model flags uncertainty and triggers retrieval to resolve the ambiguity, correctly identifying it as a Title. Conversely, in the confident Case 2 ("Steven

Spielberg"), the model detects a clear intent pattern and performs direct prediction without retrieval overhead. This comparison indicates that retrieval decisions are driven by introspective judgment rather than heuristics: the model seeks external knowledge only when internal evidence is lacking, bypassing retrieval for simple cases to balance efficiency and accuracy. This validates that NE-R1 has effectively learned an intuitive "retrieval-on-demand" strategy. The corresponding visualized trajectories are shown in Figure 10.

False-positive and false-negative retrieval cases.

Beyond successful examples, we also inspect failure modes of the learned retrieval policy. Figure 11 reports one false-positive retrieval case, where the model retrieves despite sufficient local evidence, and one false-negative case, where the model fails to retrieve for an ambiguous or knowledge-intensive mention. These cases clarify the remaining limitations of on-demand retrieval.

MIT-Movie Task 2: Retrieval Triggering (T_{trigger})

You are a senior NER expert. Determine whether search is necessary before MIT-Movie NER.

Search Rules: (1) Search when the context cannot determine whether a span belongs to an entity category. (2) Search when a mention is ambiguous, e.g., it may refer to an actor, director, character, common noun, or movie title. (3) Do not search when the intent and entity type are clear from context.

Examples:

Query: matrix

Thought: "matrix" may be a movie title; search can verify it.

Search Query: matrix movie

Query: 3d

Thought: In the movie domain, "3d" clearly refers to a genre or format; no search is needed.

Search Query:

Query: what is brave about

Thought: "Brave" may be a movie title; search can verify it.

Search Query: Brave movie

Query: did ray liotta direct any films

Thought: The mention is clearly a person name; no search is needed for NER.

Search Query:

Output Format: <think>...</think>; <thought>...</thought>; <search_q>...</search_q>.

Start: Query: what movies star bruce willis

Thought:

Figure 7: MIT-Movie prompt for the retrieval triggering task. This task initializes the model to reason about retrieval necessity and generate search queries.

MIT-Movie Task 3: Fusion Inference (T_{rag})

You are a senior NER expert. Please extract entity spans from the given MIT-Movie query. Search results are provided as external evidence; use them to resolve ambiguous names, long-tail movie titles, actors, directors, characters, and other knowledge-intensive mentions. Do not copy irrelevant evidence into the answer.

Entity Categories: ACTOR: real-world actors' names; CHARACTER: movie characters; DIRECTOR: film directors' names; GENRE: movie genres; PLOT: storyline phrases; RATING: age-rating labels; RATINGS_AVERAGE: scores, star ratings, or recommendation levels; REVIEW: subjective review terms; SONG: song titles; TITLE: official movie titles; TRAILER: trailer or preview mentions; YEAR: years or time expressions related to movies.

Output Rules: (1) Extract spans from the original query and keep their original surface form. (2) Return a JSON string whose keys are entity categories and whose values are span lists. (3) Omit empty categories. (4) Use retrieved evidence only to determine the entity type; the final span must come from the query rather than the retrieved snippets.

Retrieved Evidence: WebSearch(key="Harold Ramis actor Bill Murray actor"): 1. Harold Ramis is a director, actor, and writer. Representative works include Groundhog Day. 2. Bill Murray is an actor and comedian. Representative works include Groundhog Day and Ghostbusters. 3. Groundhog Day cast information lists Harold Ramis and Bill Murray.

Start Annotating: Query: find me all of the movies that starred harold ramis and bill murray

Output: {"ACTOR": ["harold ramis", "bill murray"]}

Figure 8: MIT-Movie prompt for the fusion inference task. Retrieved evidence is provided to help the model resolve ambiguous or knowledge-intensive mentions before producing the final NER output.

MIT-Movie Evaluation Prompt with Adaptive Retrieval

You are a senior NER expert. Please extract entity spans from the given MIT-Movie query. You may either answer directly or call a search engine when external knowledge is needed.

Entity Categories: ACTOR: real-world actors' names; CHARACTER: characters in movies; DIRECTOR: film directors' names; GENRE: movie genres, such as action, comedy, horror, and 3D; PLOT: phrases describing the storyline, such as vampires, ghosts, love at first sight, war, or explosions; RATING: age-rating labels, e.g., pg 13 or nc 17; RATINGS_AVERAGE: scores, star ratings, or recommendation levels; REVIEW: subjective review terms; SONG: song titles; TITLE: official movie titles; TRAILER: trailer or preview mentions; YEAR: years or time expressions related to movies.

Extraction Rules: (1) Extract spans based on query context and keep the original text. (2) Return a JSON string whose keys are entity categories and whose values are span lists. (3) Omit empty categories. (4) If search is used, use retrieved information only as evidence for disambiguation; do not output entities that do not appear in the query.

Search Rules: Call search when the context cannot determine whether a span belongs to one of the entity categories, or when a mention is ambiguous, such as a name that may refer to an actor, director, character, common noun, or movie title. Do not search if the query semantics are already clear.

Output Format: Need to search: <search>search query</search> <information>search info...</information> <answer>final answer</answer>. No need to search: <answer>final answer</answer>. The final answer must be valid JSON inside the answer tag.

Start Annotating: Query: {user_input}

Figure 9: MIT-Movie evaluation prompt with adaptive retrieval. The model either triggers search and fuses retrieved information, or directly returns the final NER answer.

Case 1: Ambiguity Resolution (Hard Sample)

Input : “I am looking for information about Crash.”

Model Inference:

- <think> : The span “Crash” is ambiguous. It can refer to a common noun (accident) or a movie title... the type (TITLE vs non-entity) is unclear. Context alone is insufficient.</think> **[Decision:SEARCH]**
- <search>Crash movie information</search>
- <information>Search results show that “Crash” is a well-known movie title, including: Crash (1996), directed by David Cronenberg

Output : <answer>(Crash, TITLE)</answer>

Case 2 : Confident Prediction (Easy Sample)

Input : “find movies directed by steven spielberg”

Model Inference:

- <think> : The query follows a standard movie search intent pattern. “steven spielberg” matches a director name commonly annotated in the dataset.</think> **[Decision:DIRECT]**

Output : <answer>(steven spielberg, DIRECT)</answer>

Figure 10: Different handling strategies for simple samples and difficult samples in NE-R1

False Positive Retrieval: unnecessary search on an easy instance

Input: “find action movies with a lot of explosions”

Model trajectory: <think> The query mentions “explosions”, which might refer to specific movie titles. External information may help identify relevant titles. </think> **[Decision: SEARCH]** <search>explosion action movies</search>

Retrieved information: “Top action movies featuring explosions: Die Hard, Speed, Mission Impossible...”

Prediction: {"Genre": ["action"], "Plot": ["explosions"]}

Ground truth: {"Genre": ["action"], "Plot": ["explosions"]}

Analysis: The retrieved results introduce many concrete movie titles, but the input only requires recognizing the genre and plot description. Retrieval is therefore redundant and may introduce noise, although it does not change the final prediction in this case.

False Negative Retrieval: missed search on a hard instance

Input: “I want to watch something like primer”

Model trajectory: <think> The word “primer” could refer to a common noun, such as a basic guide. The context is insufficient to confirm it as a movie title. </think> **[Decision: DIRECT]**

Prediction: misses the title entity “primer”.

Ground truth: {"Title": ["primer"]}

Analysis: “Primer” is a long-tail entity referring to a 2004 low-budget independent science-fiction film. Since the model’s parametric knowledge is insufficient to confidently identify it as a movie title, retrieval should have been triggered. The missed search leads to an entity-type recognition error.

Figure 11: Representative false-positive and false-negative retrieval cases on MIT-Movie. The false-positive case shows redundant retrieval on an easy query, while the false-negative case shows a missed retrieval opportunity for a long-tail movie title.