SAM-NER: Semantic Archetype Mediation for Zero-Shot Named Entity Recognition

Ruichu Cai[1,2], Juntao Gan[1], Miao Mai[3], Zhifeng Hao[1,4], Boyan Xu[1]*
[1]School of Computer Science, Guangdong University of Technology
[2]Peng Cheng Laboratory [3]Nanfang Media Group(Nanfang Daily)
[4]College of Mathematics and Computer, Shantou University
{cairuichu, yinghuo.gan}@gmail.com maim@nfmedia.com
haozhifeng@stu.edu.cn hpakyim@gmail.com

Abstract

Zero-shot Named Entity Recognition (ZS-NER) remains brittle under domain and schema shifts, where unseen label definitions often misalign with a large language model's (LLM's) intrinsic semantic organization. As a result, directly mapping entity mentions to fine-grained target labels can induce systematic semantic drift, especially when target schemas are novel or semantically overlapping. We propose **SAM-NER**, a three-stage framework based on *Semantic Archetype Mediation* that stabilizes cross-domain transfer through an intermediate, domain-invariant archetype space. SAM-NER: (i) performs *Entity Discovery* via cooperative extraction and consensus-based denoising to obtain high-coverage, high-fidelity entity spans; (ii) conducts *Abstract Mediation* by projecting entities into a compact set of universal semantic archetypes distilled from high-level ontological abstractions; and (iii) applies *Semantic Calibration* to resolve archetype-level predictions into target-domain types through constrained, definition-aligned inference with a frozen LLM. Experiments on the CrossNER benchmark show that SAM-NER consistently outperforms strong prior ZS-NER baselines in cross-domain settings.

1 Introduction

Zero-shot Named Entity Recognition (ZS-NER) aims to identify and type entities in unseen domains or under novel label taxonomies without target-domain supervision (Wei et al., 2024; Xie et al., 2024). With their strong general language understanding and broad world knowledge, large language models (LLMs) have become the dominant paradigm for tackling ZS-NER, enabling flexible reasoning over natural language descriptions of entity types and extraction rules. Most existing LLM-based ZS-NER approaches can be broadly categorized into two paradigms: (i) instruction-based structured extraction via natural language constraints, and (ii) retrieval-augmented inference (RAG) that incorporates external evidence during generation or verification.

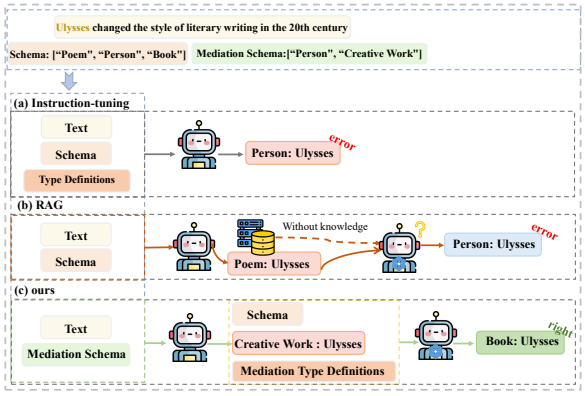

Figure 1: Comparison of different ZS-NER methods

Instruction-based methods reformulate entity specifications and task constraints into prompts that guide LLMs to perform span extraction and type assignment. For example, Sainz et al. (2024) maps structured task specifications into natural-language annotation instructions, allowing LLMs to follow explicit extraction principles across diverse domains. In contrast, retrieval-augmented approaches enhance inference by incorporating exogenous knowledge sources; for instance, Kim et al. (2024) retrieves background evidence to support post-hoc verification and error correction, aiming to reduce hallucinations and improve reliability. Both paradigms have demonstrated promising performance under zero-shot and low-resource settings, highlighting the potential of LLMs as a unifying backbone for NER.

However, existing LLM-based ZS-NER methods exhibit fundamental limitations when applied to cross-domain settings with heterogeneous label semantics. Instruction-based approaches im-

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 41216–41231
July 2-7, 2026

plicitly assume that target-domain label definitions align well with the model's intrinsic semantic organization—an assumption that often fails when fine-grained, domain-specific schemas introduce systematic semantic drift. Meanwhile, retrieval-augmented methods are inherently constrained by the availability, coverage, and reliability of external knowledge sources, which are frequently sparse or incomplete in specialized vertical domains. As illustrated in Figure 1(a) and (b), both forms of mismatch—semantic misalignment between model representations and target definitions, and insufficiencies in external knowledge—lead to characteristic failure modes in zero-shot extraction, resulting in distorted or erroneous entity predictions. These observations motivate the need for an intermediate semantic abstraction that can stabilize cross-domain transfer without relying on domain-specific supervision or external knowledge.

To address these challenges, we propose **SAM-NER**, a zero-shot NER framework based on the principle of *Semantic Archetype Mediation*. Our key insight is that while target-domain label taxonomies are often volatile and domain-specific, the underlying semantic archetypes instantiated by entities remain largely invariant across domains. Instead of forcing LLMs to directly map entity mentions to narrow and unseen target labels, SAM-NER introduces an intermediate mediation layer that stabilizes semantic reasoning by decoupling semantic understanding from schema-specific label definitions. SAM-NER operates as a three-stage pipeline. (i) *Entity Discovery via Cooperative Extraction* identifies a high-recall yet high-fidelity set of candidate entity spans through dual-source extraction and consensus-based denoising; (ii) *Abstract Mediation via Semantic Archetypes* projects these candidates into a universal archetype space, establishing stable and domain-invariant semantic anchors; and (iii) *Definition-Guided Semantic Calibration* resolves archetype-level predictions into fine-grained target-domain types through constrained, definition-aligned inference with a frozen LLM. As illustrated in Figure 1(c), this mediation-driven process enables reliable discrimination between semantically adjacent target types even under severe schema shift.

The main contributions of this work are summarized as follows:

- We introduce *Semantic Archetype Mediation*, a new paradigm for zero-shot NER that stabilizes cross-domain generalization by decoupling semantic understanding from volatile label definitions.
- We propose a cooperative entity discovery framework that combines a precision-oriented anchor extractor with a recall-oriented explorer extractor, and introduce consensus-based denoising to reconcile their complementary error profiles.
- We develop a definition-guided semantic calibration mechanism that grounds archetype-level predictions into target-domain labels through constrained, definition-aligned reasoning, eliminating reliance on external knowledge bases.
- Extensive experiments on CrossNER demonstrate that SAM-NER achieves state-of-the-art performance, validating semantic archetype mediation as a more robust alternative to direct label mapping for out-of-domain generalization.

2 Related Work

2.1 Generative Models for NER

Instruction tuning is a paradigm that was widely adopted in NER during the early stages of large language model development. Prior work injects task priors and entity type semantics into LLMs through natural language instructions, such as annotation guidelines, extraction rules, and type descriptions, which has been shown to effectively improve entity boundary detection and type discrimination, particularly in zero-shot and low resource settings (Wang et al., 2023; Zhou et al., 2023; Bai et al., 2024; Keloth et al., 2024).

To further enhance the stability of instruction representations, recent studies introduce more structured formulations. Approaches such as Code4UIE (Guo et al., 2024) and KnowCoder (Li et al., 2024) encode extraction schemas and type constraints as executable code or structured class hierarchies, explicitly modeling semantic relations among entity types while retaining the core instruction following paradigm.

In addition, with the further development of large language models, agent-based methods have emerged as a significant direction in NER research. This approach decomposes NER into multiple interactive agents and enhances robustness in zero-shot

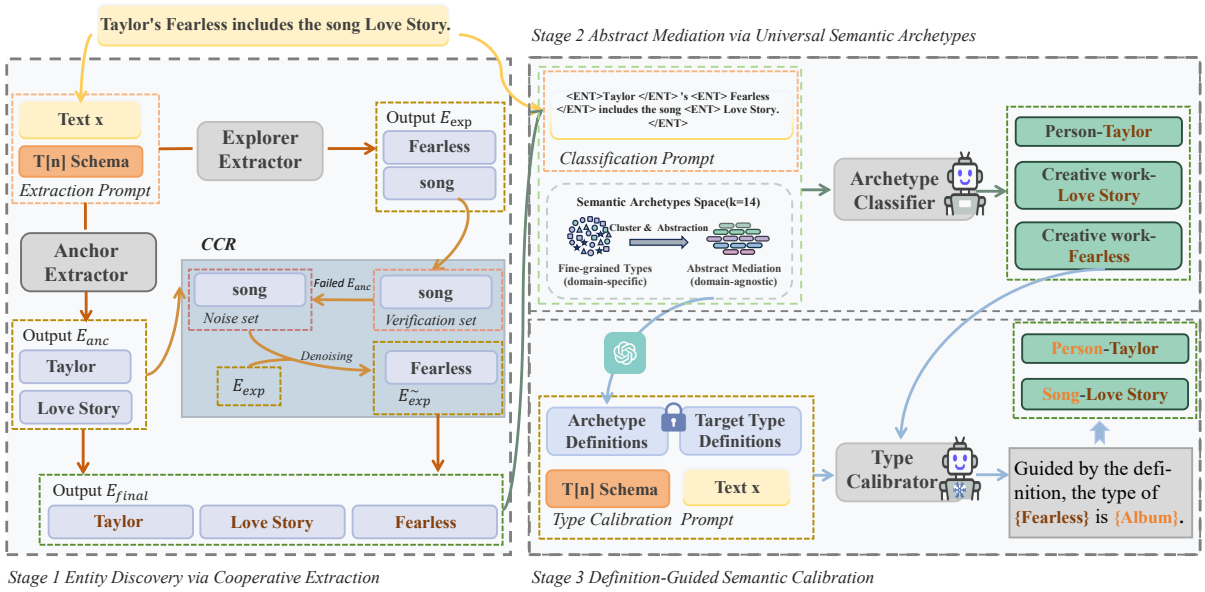

Figure 2: **The overall framework**. In Stage 1, extracted entities are used to annotate sentences. In Stage 2, archetype mediation are assigned to the marked entities. In Stage 3, the mediation types are refined into target domain types based on type definitions.

or low-resource scenarios through structured interactions, tool usage, or iterative improvements. Key mechanisms include: self-iteration/reflection, which involves reviewing and correcting predictions or pseudo-labels to refine entity boundaries and types (Wang et al., 2025; Tao et al., 2025); and networked retrieval, which utilizes an independent Knowledge Retrieval Agent to obtain external knowledge from Wikipedia to enrich contextual information (Mu et al., 2026).

Despite their effectiveness, instruction tuning methods generally assume that LLM internal semantic representations are well aligned with target domain entity type semantics. This assumption often breaks down in out-of-domain or novel type scenarios, where semantic misalignment limits generalization performance. In contrast, agent-based methods leverage multi-step collaboration among different agents to enhance the accuracy of reasoning, with performance improvements primarily stemming from external coordination and multi-step collaboration.

2.2 Entity Type Descriptions without External Knowledge

In knowledge intensive Named Entity Recognition scenarios, retrieval-augmented generation (RAG) has been widely adopted to enhance entity recognition and type discrimination. Such approaches typically assist LLMs by retrieving similar annotated examples from demonstration pools or by incorporating entity-related background knowledge from external knowledge bases during inference (Nandi and Agrawal, 2024; Liao et al., 2025; Cai et al., 2025; Keloth et al., 2024). Although these methods achieve substantial performance gains across multiple benchmarks, their effectiveness heavily depends on the quality and coverage of external knowledge resources, limiting their scalability to data-scarce domains.

In the absence of external knowledge, recent studies have explored leveraging natural language descriptions—such as entity type definitions—to provide semantic constraints for NER (Sainz et al., 2024; Zamai et al., 2024; Cocchieri et al., 2025a,b). Instead of treating entity types as discrete symbolic labels, these methods convert them into human-readable forms, including type definitions, annotation guidelines, or labeling instructions, and incorporate them as soft constraints during extraction and classification. This formulation enables LLMs to perform entity typing based on semantic understanding, even without explicit domain knowledge.

Nevertheless, these approaches implicitly rely on a stable alignment between target type semantics and the model's intrinsic semantic space. Under

domain shift or unseen label settings, directly reasoning over target type definitions can introduce semantic mismatch, leading to type ambiguity or unstable predictions. This limitation motivates further investigation into how type semantics can be leveraged more robustly for out-of-domain NER without external knowledge.

3 Methodology

SAM-NER is a three-stage framework designed to mitigate semantic drift in zero-shot Named Entity Recognition through *semantic archetype mediation*. As illustrated in Figure 2, the framework follows a progressive mediation pipeline that transitions from entity span discovery to abstract semantic anchoring, and finally to definition-grounded target typing. Specifically, (i) the **Entity Discovery** stage via Cooperative Extraction identifies a high-recall yet high-fidelity set of candidate entity spans through dual-source extraction and consensus-based denoising; (ii) the **Abstract Mediation** stage via Universal Semantic Archetypes projects these candidates into a universal archetype space, establishing stable and domain-invariant semantic anchors; and (iii) the Definition-Guided **Semantic Calibration** stage resolves archetype-level predictions into fine-grained target-domain types through constrained, definition-aligned inference. By decoupling entity discovery, semantic mediation, and target-domain grounding, SAM-NER provides a robust and interpretable solution for zero-shot NER under heterogeneous label semantics.

3.1 Entity Discovery via Cooperative Extraction

This stage aims to construct a high-coverage yet high-fidelity set of *type-agnostic* entity span candidates for subsequent semantic archetype mediation. We adopt a dual-source design to exploit complementary error profiles: a precision-oriented *anchor extractor* provides stable corroboration signals, while a silver-supervised *explorer extractor* improves recall but may introduce spurious mentions. We then perform consensus-based denoising to suppress silver-noise without sacrificing coverage.

Anchor Extractor. We instantiate the anchor extractor $Extractor_{\text{anc}}$ as a precision-oriented instruction-tuned model. Specifically, we use Llama3-8B-Instruct fine-tuned on high-quality IE instructions from IEPile (Gui et al., 2024), which exhibits strong boundary discrimination and stable semantic behavior under domain shift. Given an input sentence x_i and a target type set T_i, the anchor extractor outputs a set of candidate entity mentions:

$$E^i_{\text{anc}} = Extractor_{\text{anc}}(x_i, T_i). \tag{1}$$

Each $e \in E^i_{\text{anc}}$ denotes an extracted entity span in x_i.

Explorer Extractor. To increase candidate coverage and capture entities that may be missed by the anchor extractor, we introduce a recall-oriented explorer extractor $Extractor_{\text{exp}}$ trained with broad-spectrum silver supervision. We train $Extractor_{\text{exp}}$ using Pile-NER (Zhou et al., 2023) by converting its dialogue-style supervision into a standard instruction-following format. Given the same input (x_i, T_i), the explorer extractor produces an additional candidate set:

$$E^i_{\text{exp}} = Extractor_{\text{exp}}(x_i, T_i). \tag{2}$$

In practice, $Extractor_{\text{exp}}$ tends to over-generate low-salience, word-level mentions (e.g., "data", "system", "user"). We attribute this behavior to noise in the **silver-standard annotations** of Pile-NER, where automated labeling may spuriously mark generic functional nouns as entities, particularly under domain shift. Such spurious candidates introduce substantial variance and motivate an explicit denoising mechanism before archetype mediation.

Collaborative Consensus Refinement. We introduce Collaborative Consensus Refinement (CCR) as a consensus-based denoising mechanism to refine the explorer candidates produced by the silver-supervised extractor. CCR leverages the anchor extractor as an independent semantic validator and selectively filters noise-prone candidates in E^i_{exp} through cross-model consensus. The underlying intuition is that generic, non-referential spans are prone to appear in silver-supervised outputs, whereas the instruction-tuned anchor extractor provides a stronger semantic prior for validating whether a span constitutes a meaningful entity mention in context.

Concretely, we identify a subset of explorer candidates that are most prone to silver-noise artifacts and aggregate them into a verification set $V^i_{\text{noise}} \subseteq E^i_{\text{exp}}$. A candidate $e \in V^i_{\text{noise}}$ is marked as

noise only if it fails to obtain independent corroboration from the anchor extractor:

$$D^i_{\text{noise}} = \{\, e \in V^i_{\text{noise}} \mid e \notin E^i_{\text{anc}} \,\}. \tag{3}$$

The denoised explorer candidate set is then obtained as:

$$\tilde{E}^i_{\text{exp}} = E^i_{\text{exp}} \setminus D^i_{\text{noise}}. \tag{4}$$

Finally, we construct the discovered entity set via a de-duplicated union:

$$E^i_{\text{final}} = E^i_{\text{anc}} \cup \tilde{E}^i_{\text{exp}}. \tag{5}$$

This consensus-driven refinement preserves the high-precision predictions of the anchor extractor while retaining the coverage benefits of the explorer extractor, yielding a high-fidelity candidate set for subsequent semantic archetype mediation.

Algorithm 1: Collaborative Consensus Refinement (CCR)

Input: E_{anc}, E_{exp}
$D_{noise} \leftarrow \emptyset,\ \tilde{E}^i_{\text{exp}} \leftarrow \emptyset$
for $e^i_{exp} \in E_{exp}$ **do**
  $len_e \leftarrow len(\text{split}(e^i_{exp}))$
  if $len_e = 1$ & $e^i_{exp} \notin E_{anc}$ **then**
    $D_{noise} \leftarrow D_{noise} \cup \{e^i_{exp}\}$
  end if
end for
for $e^i_{exp} \in E_{exp}$ **do**
  if $e^i_{exp} \notin D_{noise}$ **then**
    $\tilde{E}^i_{\text{exp}} \leftarrow \tilde{E}^i_{\text{exp}} \cup \{e^i_{exp}\}$
  end if
end for
Output: $\tilde{E}^i_{\text{exp}}$

3.2 Abstract Mediation via Universal Semantic Archetypes

The core bottleneck in zero-shot Named Entity Recognition lies in the high variance and domain-specificity of target label semantics.

To address this challenge, we introduce an intermediate *Abstract Mediation* stage—*Semantic Archetype Mediation*—that bridges entity discovery and target-domain calibration. Our key hypothesis is that while fine-grained entity labels vary substantially across domains and annotation schemas, they can be projected into a compact and stable semantic space consisting of broad, invariant archetypes. By performing classification within this archetype space, we decouple semantic understanding from volatile label definitions and reduce the alignment pressure between the LLM's intrinsic representations and heterogeneous domain-specific schemas.

Ontological Distillation and Archetype Mapping. We construct the mediation space by consolidating heterogeneous NER schemas from IEPile (Gui et al., 2024), which integrates multiple information extraction benchmarks with instance-specific type sets. Let D_{ner} denote the NER subset of IEPile. For each training instance i, let $\mathcal{T}^i_{\text{orig}}$ denote its original fine-grained type schema, which may differ substantially across data sources.

Guided by taxonomic design principles proposed by Yang et al. (2025), we distill a universal set of 14 semantic archetypes $\mathcal{A}$ (e.g., *Person*, *Medicine*) to serve as the shared mediation space for zero-shot NER. Rather than adopting an existing ontology verbatim, we construct an abstract type system explicitly tailored for semantic mediation across heterogeneous NER schemas. The choice of archetype granularity is empirically motivated and further analyzed in Section 6.

The resulting archetypes are designed to satisfy several desiderata: (i) *readability*, ensuring that type names are interpretable to both humans and language models; (ii) *unambiguity*, minimizing semantic overlap between closely related categories; (iii) *hierarchical coherence*, aligning fine-grained types with their abstract parents; and (iv) *flexibility*, allowing the taxonomy to accommodate diverse NER tasks and emerging entity types.

We formalize schema consolidation via a deterministic projection function:

$$M : \mathcal{T}_{\text{orig}} \to \mathcal{A}, \tag{6}$$

where $\mathcal{T}_{\text{orig}} = \bigcup_i \mathcal{T}^i_{\text{orig}}$ denotes the union of fine-grained types observed in D_{ner}. For each instance i, the corresponding abstract schema is obtained by projecting its original schema:

$$\mathcal{A}^i = M(\mathcal{T}^i_{\text{orig}}) \subseteq \mathcal{A}. \tag{7}$$

The projection function M is constructed through a data-informed, principle-guided schema design process. Semantically related fine-grained labels are grouped under shared abstract parents according to their conceptual scope and hierarchical relations, while enforcing the desiderata above. Importantly, M is deterministic and fixed throughout training and inference. For reproducibility, the

complete type-to-archetype mapping is provided in Appendix B.

Applying M to the original annotations resolves ontological overlaps and homogenizes the label space. To preserve the relationship between each entity mention and its sentential context, we employ an entity-aware formatting strategy that marks mentions with <ENT> and </ENT> tags. The resulting abstract training set is defined as:

$$D_{\text{abs}} = \{(S^i_{\text{tag}}, \mathcal{A}^i, Y^i_{\text{abs}})\}_{i=1}^{|D_{\text{ner}}|}, \tag{8}$$

where S^i_{tag} is the tagged sentence and Y^i_{abs} denotes the archetype-level annotations obtained by projecting the original labels through M. This design allows the classifier to respect instance-specific schema constraints during training, while learning a unified and domain-invariant abstraction space.

Abstract Archetype Classifier. We train an abstract archetype classifier $Classifier_{\text{abs}}$ on D_{abs} to serve as the mediation module. During training, the classifier is conditioned on the instance-specific abstract schema $\mathcal{A}^i$, reflecting the multi-source setting where the candidate type set varies across examples. During inference, since the input consists only of entity-marked sentences without schema constraints, the classifier predicts over the full archetype space $\mathcal{A}$:

$$y^i_{\text{abs}} = Classifier_{\text{abs}}(x^i_{\text{tag}}, \mathcal{A}), \tag{9}$$

where y^i_{abs} represents the predicted archetype assignments for entities in the sentence. The collection $Y_{\text{abs}} = \{y^i_{\text{abs}}\}_{i=1}^n$ provides stable semantic anchors for the subsequent definition-guided calibration stage, enabling target-domain typing to be grounded in a pre-aligned and domain-invariant semantic framework.

3.3 Definition-Guided Semantic Calibration

To bridge the final gap between the abstract archetype space $\mathcal{A}$ and domain-specific taxonomies, we introduce a *Definition-Guided Semantic Calibration* stage. Given the archetype predictions Y_{abs} produced by the semantic archetype mediation stage, this module resolves each archetype assignment into a fine-grained target-domain label through *constrained type inference*. Unlike approaches that directly prompt LLMs to predict target labels, we leverage archetypes as semantic anchors and explicitly condition target-type reasoning on the predicted y^i_{abs}, thereby narrowing the inference space and mitigating semantic drift in zero-shot transfer.

Axiomatic Definition Construction. To establish clear and stable semantic boundaries for the archetype space, we associate each abstract archetype $a \in \mathcal{A}$ with a canonical semantic definition. Let

$$\mathcal{D}_{\text{abs}} = \{d_a \mid a \in \mathcal{A}\} \tag{10}$$

denote the set of abstract type definitions, where each d_a characterizes the essential semantic scope of archetype a. These definitions are synthesized via prompt-based distillation using a large language model (OpenAI, 2025) and are designed to be stylistically consistent and semantically exclusive. As a result, they serve as explicit semantic constraints that delineate the abstract mediation space without introducing domain-specific leakage.

For a given target domain, let $\mathcal{T}_{\text{tgt}}$ denote its entity type set. We associate each target type $t \in \mathcal{T}_{\text{tgt}}$ with a refined natural language definition and denote the resulting definition set as:

$$\mathcal{D}_{\text{tgt}} = \{d_t \mid t \in \mathcal{T}_{\text{tgt}}\}. \tag{11}$$

We apply minimal linguistic normalization to the original benchmark descriptions (e.g., CrossNER (Liu et al., 2021)) to remove syntactic redundancies and underspecified semantics while preserving their original intent. These refined target definitions constitute the axiomatic constraints for definition-guided calibration.

Constrained Definition-Aligned Inference. We employ a frozen large language model as a calibration operator, denoted by $Calibrator$, to perform definition-aligned type inference. For each extracted entity instance i, we are given its context sentence S^i, its predicted abstract archetype $y^i_{\text{abs}} \in \mathcal{A}$, and the target-domain type definitions $\mathcal{D}_{\text{tgt}}$. The calibration objective is to resolve the most compatible target type under the constraint imposed by the archetype prior:

$$y^i_{\text{tgt}} = \arg\max_{t \in \mathcal{T}_{\text{tgt}}} \text{Align}\big(d_{y^i_{\text{abs}}}, d_t \mid S^i\big), \tag{12}$$

where $d_{y^i_{\text{abs}}} \in \mathcal{D}_{\text{abs}}$ is the definition of the predicted archetype, and $\text{Align}(\cdot)$ denotes a definition-alignment scoring function implicitly realized by the frozen LLM. This function evaluates the semantic entailment and compatibility between the archetype prior and candidate target definitions within the given context.

By enforcing this abstract-to-target alignment constraint, the calibrator avoids unconstrained label reasoning and instead grounds target-type selection in a pre-stabilized semantic framework. This dual-level alignment enables robust discrimination between semantically adjacent target types, even when they are unseen during training.

Finally, the calibrated predictions across all extracted entities are aggregated as:

$$Y_{\text{tgt}} = \{y^i_{\text{tgt}}\}_{i=1}^{N}, \tag{13}$$

where N is the number of extracted entity instances. This calibrated entity set constitutes the final output of **SAM-NER**.

4 Experimental Settings

4.1 Datasets

Training Sets. In our experiments, we use Pile-NER (Zhou et al., 2023) and IEPile (Gui et al., 2024) as training datasets. Pile-NER covers approximately 13K entity types and contains around 240K entity instances, exhibiting a highly diverse distribution of entity types. IEPile is an instruction-tuning dataset for information extraction on a large scale that integrates 33 widely used IE benchmarks, in this work, we use only its Named Entity Recognition subset.

Benchmark. We adopt CrossNER (Liu et al., 2021) as the evaluation benchmark to assess the effectiveness of the proposed approach on out-of-domain zero-shot Named Entity Recognition. CrossNER spans multiple domains, including Artificial Intelligence, Literature, Music, Politics, and Science, enabling systematic evaluation of model generalization across diverse domains.

4.2 Evaluation Metrics

We use micro-F1, a widely adopted metric in Named Entity Recognition, as the evaluation measure.

4.3 Compared Baselines

We select the following representative zero-shot and out-of-domain Named Entity Recognition methods as baselines for comparison:

- **InstructUIE** (Wang et al., 2023) InstructUIE jointly instruction tuning multiple IE tasks under a unified framework.
- **UniNER** (Zhou et al., 2023) UniNER combines target distillation with instruction tuning for cross-task generalization.
- **IEPile** (Gui et al., 2024) This method constructs a large data set to train the information extraction ability of the model.
- **GoLLIE** (Sainz et al., 2024) GoLLIE trains LLMs with structured annotation guidelines for type-driven extraction.
- **KnowCoder** (Li et al., 2024) KnowCoder encodes structured constraints as executable code to enhance semantic understanding.
- **GLiNER** (Zaratiana et al., 2024) a lightweight general-purpose named entity recognition model based on bidirectional Transformer, which performs exceptionally well in resource-constrained scenarios.
- **IRRA** (Cai et al., 2025) IRRA applies retrieval augmented generation to refine entity typing. Since the optimal settings introduce additional knowledge, but the Guidelines settings are similar to ours, we chose to use the Guidelines settings for comparison.
- **GUIDEX** (Fuente et al., 2025) GUIDEX performs data synthesis guided by the schema and infers entities using type definitions.

For all baselines, we report the best results under the zero-shot and out-of-domain settings as documented in the original papers.

4.4 Backbones & Implementation

In all experiments, we consistently used Llama3-8B-Instruct as the backbone model. Additionally, to validate the generalization ability of the proposed method across different model families, we further introduced Qwen2.5-7B-Instruct as a supplementary backbone model. It should be noted that since the Anchor Extractor utilizes the Llama3-8B LoRA weights provided by IEPile(Gui et al., 2024), Qwen2.5-7B-Instruct is used only as the base model for the Explorer Extractor, Archetype Classifier, and Type Calibrator in this setup. Both the extractors and the classifier are trained via supervised instruction tuning, with fine-tuning that updates only a small subset of parameters using LoRA (Hu et al., 2022). All models are trained on three NVIDIA RTX 3090 GPU using the LlamaFactory framework (Zheng et al., 2024).

Method	Params	Backbone	AI	Literature	Music	Politics	Science	Avg.
InstructUIE (2023)	11B	Flan-T5	49.0	42.7	53.2	48.1	49.2	48.4
UniNER (2023)	13B	LLaMA	54.2	60.9	64.5	61.4	63.5	60.9
IEPile-Llama3-8B (2024)	8B	Llama3	50.2	43.3	53.7	57.0	50.4	50.9
GoLLIE (2024)	34B	CodeLLaMA	61.6	59.1	68.4	60.2	56.3	61.1
KnowCoder (2024)	7B	LLaMA2	60.3	61.1	70.0	72.2	59.1	64.5
GLiNER-Large (2024)	0.3B	DeBERTa-v3	57.2	64.4	69.6	**72.6**	62.6	65.3
IRRA-Guidelines (2025)	8B	LLaMA3	53.2	57.7	64.7	66.2	64.1	61.2
GUIDEX (2025)	8B	LLaMA3.1	**62.4**	63.8	67.9	69.6	64.6	65.7
SAM-NER(Qwen2.5-7B)	7B	Qwen2.5	57.9	64.1	69.3	66.7	62.1	64.3
SAM-NER(Llama3-8B)	8B	LLaMA3	58.2	**68.7**	**71.2**	68.2	**65.1**	**66.3**

Table 1: The micro-F1 scores on zero-shot cross-domain setting. Except for IRRA, all baseline scores are directly taken from the results reported in the original papers under their optimal settings. We used the scores generated by IRRA under the Guidelines settings for comparison.

5 Results

Table 1 presents a comparison between our SAM-NER and several state-of-the-art zero-shot NER systems. When using Llama3-8B as the backbone, our method achieved an average F1 score of 66.3 on the CrossNER, outperforming all comparison methods in the benchmark. Specifically, SAM-NER achieved the best performance in the literature, music, and science domains, with a 4.3-point increase in F1 score over the second-best method in the literature domain. When using Qwen2.5-7B as the backbone, the model's average F1 score also approached the runner-up's level, further demonstrating the effectiveness of the proposed abstract mediation mechanism in stabilizing cross-domain transfer.

We observe relatively lower performance in the AI; However, UniNER and IEPile-Llama3-8B, trained on the same data as our Cooperative extractor, exhibit similar trends, suggesting that this problem may stem from an intrinsic bias in the training data. GLiNER's performance advantages in the political domain may stem from the stability of its entity structure and clear semantic boundaries, characteristics that align highly effectively with its entity span-type semantic matching mechanism. When the semantic distance between target domain types and the abstract mediation layer is insufficient, the prototype's mediating role may become less effective.

6 Analysis

Contribution of Different Components. To demonstrate the contributions of the entity discovery via cooperative extraction and definition-guided semantic calibration stages to SAM-NER under the ZS-NER setting, we conducted further analysis. During the entity discovery via cooperative extraction phase, we removed the anchor extractor (**w/o anc.**) and the explorer extractor (**w/o exp.**) respectively. We removed the definition-guided semantic calibration stage (**w/o cali.**) and retrained the classification model on data with unmapped abstract Archetypes to adapt it for predicting target entity types.

Dataset	w/o exp.	w/o anc.	w/o cali.
AI	53.0$_{-5.2}$	54.1$_{-4.1}$	48.5$_{-9.7}$
Literature	64.6$_{-4.1}$	61.9$_{-6.8}$	56.1$_{-12.6}$
Music	65.3$_{-5.9}$	66.1$_{-5.1}$	58.6$_{-12.6}$
Politics	63.6$_{-4.6}$	61.9$_{-6.3}$	63.7$_{-4.5}$
Science	61.4$_{-3.7}$	58.8$_{-6.3}$	54.1$_{-11}$

Table 2: Micro-F1 Values for Different Components in SAM-NER

The results are shown in Table 2. Can be observed that: (1) Removing either extractor (w/o anc. or w/o exp.) consistently leads to performance degradation. This is mainly because the two extractors are trained on different data distributions and supervision signals, resulting in divergent entity boundary predictions. (2) The performance drops substantially under the w/o cali. setting, which highlights the critical role of the intermediate semantics in out-of-domain generalization. Due to the mismatch between the intrinsic semantic representations of the language model and the target-domain type definitions, the model struggles to directly interpret target type descriptions—especially for novel types that never appear in the training data—thereby causing classification errors.

Contribution of Collaborative Consensus Refinement. Drawing on the results in Table 3 and Figure 3, the Collaborative Consensus Refinement

Method	AI	Literature	Music	Politics	Science
w/o CCR	50.8	65.3	67.2	65.5	60.9
w/ CCR	**58.2**	**68.7**	**71.2**	**68.2**	**65.1**

Table 3: Impact of the Collaborative Consensus Refinement(CCR) strategy on performance.

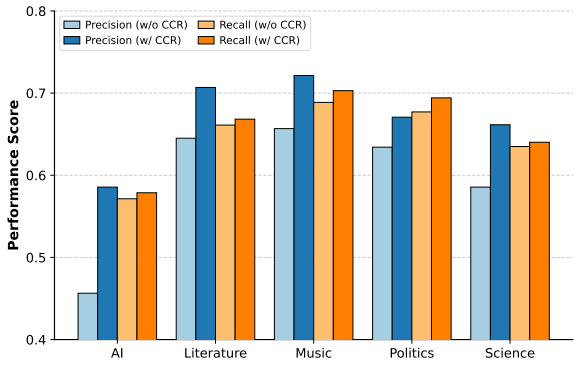

Figure 3: Impact of CCR strategy on precision and Recall in different domains

strategy effectively mitigates the spurious mention activations typical of explorer extractor trained on silver-standard data, which often misidentify generic functional terms as entity mentions. As illustrated in Table 3, the integration of CCR yields consistent F1 gains across all domains, highlighted by a 7.4-point increase in the AI domain. Metrics in Figure 3 further confirm that this improvement stems from a significant optimization in precision, successfully rectifying the precision-recall imbalance of the base extractor. By leveraging dual-model consensus as a semantic filter, CCR prunes generic lexical noise while preserving the coverage of long-tail entities. Our collaborative mechanism effectively reconciles the extensive discovery capabilities of silver-data models with the high-precision reasoning of models trained on high-quality instructions.

Why 14 Semantic Archetypes? To derive a universal abstract mediation space, we performed semantic clustering on entity patterns distilled from the IEPile corpus. In determining the optimal cluster number k, we adhered to the principle of ontological parsimony, seeking a balance between semantic granularity and cross-domain stability.

As illustrated in Figure 4(a), the Silhouette Score achieves a prominent local peak at $k = 14$. While higher numerical scores appear at $k = 21$ and $k = 24$, the Gap Statistic reveals that the structural gains from increasing k exhibit diminishing marginal returns beyond 14. Furthermore, the observed drop in the Gap Statistic at $k = 15$ signals a potential instability in semantic boundaries at that specific dimensionality. Although Figure 4(b) demonstrates that $k = 24$ can capture hyper-specific semantic nuances (e.g., "SUV" or "MPV"), such high-specificity clusters are prone to coupling with domain-specific noise in zero-shot scenarios, thereby compromising the transferability of the mediation space. Consequently, we set $k = 14$ as it represents the optimal trade-off between intra-cluster cohesion and inter-cluster separability, establishing a robust semantic space that effectively resists domain shift.

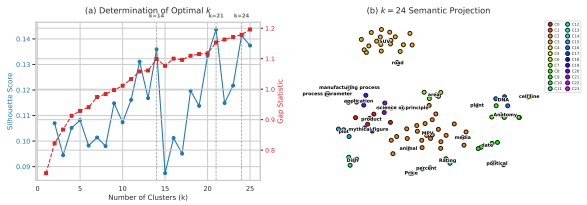

Figure 4: The Cluster Analysis. The left figure (a) shows the scores of Silhouette and Gap at different k values. The right figure (b) presents the clustering results when $k = 24$.

7 Conclusion

In this paper, we presented **SAM-NER**, a three-stage framework that mitigates semantic drift in zero-shot NER through *Semantic Archetype Mediation*. By decoupling span discovery from fine-grained target typing, SAM-NER stabilizes cross-domain transfer under heterogeneous label semantics. Concretely, we first perform *Entity Discovery* via cooperative extraction and consensus-based denoising to obtain a high-fidelity candidate set, then conduct *Abstract Mediation* by projecting candidates into a universal semantic archetype space to provide domain-invariant anchors, and finally apply *Definition-Guided Calibration* to ground archetype-level predictions into target-domain types through constrained, definition-aligned inference. Experiments on CrossNER demonstrate that SAM-NER consistently outperforms strong prior zero-shot NER baselines in cross-domain settings, with particularly pronounced gains in domains where schema-specific label semantics amplify misalignment. Overall, our findings highlight semantic archetype mediation as an effective and interpretable strategy for improving domain-invariant generalization in zero-shot NER.

Limitations

Despite its strong empirical performance, **SAM-NER** has two main limitations that warrant further investigation: **(1) Taxonomic Bias and Bounded Universality.** SAM-NER relies on a compact set of 14 semantic archetypes distilled from IEPile, which does not constitute a theoretically exhaustive ontology. As a result, the universality of the archetype space is bounded by the coverage and granularity of the distillation source. In highly specialized domains, limited semantic resolution may lead to *coarse-mapping* errors, where domain-specific nuances are absorbed into overly broad archetypes, reducing precision under fine-grained or atypical target taxonomies. **(2) Dependence on Definition Discriminability.** The fidelity of definition-guided calibration depends on the linguistic discriminability of target type definitions. When definitions are underspecified, highly overlapping, or inconsistent across labels, the constrained alignment process becomes sensitive to description quality and may produce unstable type assignments. Therefore, while semantic archetype mediation mitigates semantic drift at the schema level, the final predictions remain partly contingent on the clarity and separability of human-authored label definitions.

Acknowledgments

This research was supported in part by National Science and Technology Major Project (2021ZD0111502), Natural Science Foundation of China (U24A20233, 62406078, 62476163), the Guangdong Basic and Applied Basic Research Foundation (2023B1515120020), CCF-DiDi GAIA Collaborative Research Funds (CCF-DiDi GAIA 202521), and Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)(GML-KF-24-23).

References

Fan Bai, Junmo Kang, Gabriel Stanovsky, Dayne Freitag, Mark Dredze, and Alan Ritter. 2024. Schema-driven information extraction from heterogeneous tables. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10252–10273.

Ruichu Cai, Junhao Lu, Zhongjie Chen, Boyan Xu, and Zhifeng Hao. 2025. Handling missing entities in zero-shot named entity recognition: Integrated recall and retrieval augmentation. 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 10790–10802.

Alessio Cocchieri, Giacomo Frisoni, Marcos Martínez Galindo, Gianluca Moro, Giuseppe Tagliavini, and Francesco Candoli. 2025a. Openbioner: Lightweight open-domain biomedical named entity recognition through entity type description. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 818–837.

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

Neil De La Fuente, Oscar Sainz, Iker García-Ferrero, and Eneko Agirre. 2025. GUIDEX: Guided synthetic data generation for zero-shot information extraction. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24248–24262, Vienna, Austria. Association for Computational Linguistics.

Honghao Gui, Lin Yuan, Hongbin Ye, Ningyu Zhang, Mengshu Sun, Lei Liang, and Huajun Chen. 2024. IEPile: Unearthing large scale schema-conditioned information extraction corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 127–146, Bangkok, Thailand. Association for Computational Linguistics.

Yucan Guo, Zixuan Li, Xiaolong Jin, Yantao Liu, Yutao Zeng, Wenxuan Liu, Xiang Li, Pan Yang, Long Bai, Jiafeng Guo, and 1 others. 2024. Retrieval-augmented code generation for universal information extraction. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 30–42. Springer.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Vipina K Keloth, Yan Hu, Qianqian Xie, Xueqing Peng, Yan Wang, Andrew Zheng, Melih Selek, Kalpana Raja, Chih Hsuan Wei, Qiao Jin, and 1 others. 2024. Advancing entity recognition in biomedicine via instruction tuning of large language models. *Bioinformatics*, 40(4):btae163.

Seoyeon Kim, Kwangwook Seo, Hyungjoo Chae, Jinyoung Yeo, and Dongha Lee. 2024. VerifiNER: Verification-augmented NER via knowledge-grounded reasoning with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2441–2461, Bangkok, Thailand. Association for Computational Linguistics.

Zixuan Li, Yutao Zeng, Yuxin Zuo, Weicheng Ren, Wenxuan Liu, Miao Su, Yucan Guo, Yantao Liu, Lixiang Lixiang, Zhilei Hu, and 1 others. 2024. Knowcoder: Coding structured knowledge into llms for universal information extraction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8758–8779.

Xincheng Liao, Junwen Duan, Yixi Huang, and Jianxin Wang. 2025. Ruie: Retrieval-based unified information extraction using large language model. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9640–9655.

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*, volume 35, pages 13452–13460.

Wenxuan Mu, Jinzhong Ning, Di Zhao, and Yijia Zhang. 2026. A multi-agent llm framework for multi-domain low-resource in-context ner via knowledge retrieval, disambiguation and reflective analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 32528–32536.

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. *arXiv preprint arXiv:2411.00451*.

OpenAI. 2025. Chatgpt. https://www.openai.com/chatgpt. Accessed: 2025-06-23.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. 2024. GoLLIE: Annotation guidelines improve zero-shot information-extraction. In *The Twelfth International Conference on Learning Representations*.

Xinli Tao, Xin Dong, and Xuezhong Zhou. 2025. Oema: Ontology-enhanced multi-agent collaboration framework for zero-shot clinical named entity recognition. *arXiv preprint arXiv:2511.15211*.

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*.

Zihan Wang, Ziqi Zhao, Yougang Lyu, Zhumin Chen, Maarten de Rijke, and Zhaochun Ren. 2025. A cooperative multi-agent framework for zero-shot named entity recognition. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 4183–4195, New York, NY, USA. Association for Computing Machinery.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, and Wenjuan Han. 2024. Chatie: Zero-shot information extraction via chatting with chatgpt. *Preprint*, arXiv:2302.10205.

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, Mexico City, Mexico. Association for Computational Linguistics.

Yuming Yang, Wantong Zhao, Caishuang Huang, Junjie Ye, Xiao Wang, Huiyuan Zheng, Yang Nan, Yuran Wang, Xueying Xu, Kaixin Huang, and 1 others. 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.

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.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

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 Prompts

The full set of prompts used in SAM-NER is illustrated in Figures 5–8. These prompts decompose the NER process into four progressive stages: anchor extraction, entity exploration, archetype classification, and type calibration.

B Detailed mapping relationship in IEPile and definition of archetype

Following the empirical determination of *k=14*, we conducted a semantic consolidation of labels from the IEPile corpus to derive 14 abstract types. The systematic projection from fine-grained entity categories to these universal archetypes is summarized in Table 4. And the detailed definition of the prototype is shown in Table 5

C Complexity Analysis of SAM-NER

To evaluate the efficiency of SAM-NER, we reported the time cost and memory consumption for different module combinations. Note that all experiments were conducted on the NVIDIA A800 GPU. To objectively reflect SAM-NER's raw computational overhead, we used the Llama-Factory framework's default inference pipeline without FlashAttention, quantization, or any third-party acceleration engines, in order to provide a baseline performance reference. The results are presented in Table 6.

D Hyperparameter Settings for SAM-NER

The hyperparameters of SAM-NER are presented in Table 7.

Prompt 1: Anchor Extractor Prompt

```
{
  "instruction":  "Given a passage and a schema, your  task is to extract all named entities that conform to the
     schema type. ",

  "Schema": [List of entity types],

  "Passage": "The input sentence string"
}
```

Figure 5: Anchor Extractor Prompt. Guiding the model to identify and extract entities from text following predefined extraction schema.

Prompt 2: Explorer Extractor Prompt

```
{
  "instruction": "Please extract entities that match the schema definition from the input. Return an empty list if
     the entity type does not exist. Please respond in the format of a JSON string." ,

  "Schema": [List of entity types],

  "Input": "The input sentence string"
}
```

Figure 6: Explorer Extractor Prompt. Guiding the model to identify and extract entities from text following predefined extraction schema.

Prompt 3: Archetype Classifier Prompt

```
{
  "instruction": "Your task is to classify each entity marked with <ENT></ENT> without omission and strictly
     select the most appropriate and unique entity type from the given schema based  on the sentence
     semantics." ,

  "Schema": [List of entity types],

  "Sentence": "The input sentence string"
}
```

Figure 7: Archetype Classifier Prompt. Guiding the model to assign semantically broad types to annotated entities in the input text based on predefined abstract schema.

Prompt 4: Type Calibrator Prompt

 # Instruction
Based on the given information, you need to strictly select the appropriate target domain type from the `Candidate Types` for each named entity of the `Initial Output`.
 # Information
Text: {The input sentence string}
Initial Output: {The output of entity classifier}
Candidate Types: [List of target domain entity types]
Abstract Types Definition: {The definition of abstract types}
Target Types Definition: {The definition of target types}
 # Rules
Guidance information needed when correcting entity types.
 # Output Format

```
[{
   "entity name": "corrected type"
}]
```

Figure 8: Type Calibrator Prompt. Guiding the model to calibrate the abstract types assigned to entities to target-domain types based on both semantically broad type definitions and target-domain semantic type definitions.

Abstract Type	Fine-grained Entity Types
Person	actor, character, director, mythical figure, person
Organization	organization, media
Location	Amenity, Location, location, exact location, geographical phenomenon, geographical social political, facility, road, river, area
Biology	animal, plant, biology
Medicine	Anatomy, DNA, RNA, GENE, protein, cell line, cell type, disease, Disease, biomedical, medicine
Food	Cuisine, Dish, food, Restaurant Name, review
Vehicle	vehicle, vehicle model, vehicle range, vehicle type, vehicle velocity, brand of vehicle, color of vehicle, orientation of vehicle, position of vehicle, estate car, SUV, MPV, hatchback, roadster, sports car, sedan, coupe, trailer, van, truck, motorcycle, vintage car, bus
Creative_Work	song, work of art, title, movie, genre, creative_work
Event	event, plot
Artifact	instrument, product, artifact
Computer_Science	application, enabling technology, concept or principle, process characterization, process parameter, machine or equipment, engineering feature, machanical property, manufacturing process, manufacturing standard, computer_science
Political	law, national religious political, political
Science	astronomical object, language, material, Chemical, science
Misc	misc, else

Table 4: The detailed mapping relationship between abstract type and fine-grained entity type in IEPile.

Archetype	Archetype Definition
Person	Entities representing specific individuals identified by their proper names, including real people, fictional characters, nicknames, and aliases.
Organization	Entities representing structured groups of people working together for a common purpose, including corporations, government agencies, NGOs, musical bands, political parties, and educational institutions.
Location	Entities representing spatial or geographic regions, including countries, cities, administrative divisions, physical facilities, landmarks, and public spaces.
Biology	Entities related to living organisms and taxonomy, including animal and plant species, families, and general biological classifications.
Medicine	Entities related to healthcare and biomedical sciences, including diseases, drugs, medical procedures, anatomical structures, physiological processes, and clinical concepts.
Food	Entities related to consumables, including ingredients, prepared dishes, beverages, and culinary concepts.
Vehicle	Entities representing manufactured devices designed for transportation, including cars, aircraft, ships, spacecraft, and their specific models or classes.
Creative_Work	Entities representing distinct artistic or intellectual creations, such as books, songs, movies, video games, software titles, and media franchises.
Event	Entities representing specific occurrences or organized activities happening at a specific time and place, including festivals, wars, sports matches, conferences, and natural disasters.
Artifact	Entities representing man-made objects with specific functions, including tools, instruments, gadgets, weapons, and consumer goods.
Computer_Science	Entities related to computing and technology, including programming languages, algorithms, software architectures, technical protocols, digital metrics, and IT terminology.
Political	Entities related to governance, social structures, and ideologies, including laws, treaties, policies, religious groups, ethnicities, and sociopolitical systems.
Science	Entities related to scientific disciplines and natural phenomena, including academic fields, chemical elements, compounds, celestial bodies, and scientific theories.
Misc	Entities that cannot be clearly classified into the specific categories above, serving as a catch-all for other named entities.

Table 5: The detailed definition of archetype.

Method	Time Cost	Memory Consumption	Avg. score
w/o exp.	6486	1GPU $\times$ 29.53GB	61.6
w/o anc.	6507	1GPU $\times$ 29.53GB	60.6
w/o cali.	2883	1GPU $\times$ 17.78GB	56.2
w/ all	7247	1GPU $\times$ 29.53GB	66.3

Table 6: Complexity Analysis of SAM-NER on CrossNER. The time unit is seconds. "w/o exp." refers to without employing explorer extractor."w/o anc." refers to without using anchor extractor."w/o cali." refers to without employing definition-guided semantic calibration stage. "w/ all" refers to using all components and stages.

Hyperparameter	Explorer Extractor	Archetype Classifier
loRA_rank(r)	8	8
loRA_alpha(α)	16	16
cutoff_len	1024	1024
per_device_train_batch_size	2	2
gradient_accumulation_steps	8	8
learning_rate	2.0e-5	3.0e-5
num_train_epochs	3.0	3.0
lr_scheduler_type	cosine	cosine
warmup_ratio	0.05	0.05
dtype	bf16	bf16

Table 7: Hyperparameter Settings of SAM-NER