AN EVIDENCE-ALIGNED FRAMEWORK FOR IN-CONTEXT CHINESE LEGAL NAMED ENTITY RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Named entity recognition (NER) is a foundational information extraction task, and it must recover exact entity spans and types from limited demonstrations in the few-shot in-context settings. Chinese legal NER is particularly challenging because implicit word boundaries hinder exact span localization, while similar text spans may belong to different entity types depending on context. Existing methods often rely on whole-sequence or word-level retrieval, generate detached mention lists, creating a mismatch between the evidence provided and the evidence required for boundary- and role-sensitive decisions. In this paper, we propose **C^3-NER**, an in-context learning framework with a frozen LLM and no task-specific gradient updates. It retrieves demonstrations that contain local cues for locating entity boundaries, preserves predicted spans at their exact positions in the original text throughout extraction, and uses surrounding context to distinguish spans that may belong to different entity types. Experiments on four NER benchmarks, including complete-system and controlled-retrieval comparisons, show that C^3-NER achieves 90.99 F1 on CAIL2021 and 91.56 F1 on MSRA, outperforming the strongest reproduced baseline DEER by 3.65 and 1.87 F1 points, respectively, while ranking second on CLUENER2020 and OntoNotes. On CAIL2021, it uses one LLM call per instance and reduces LLM calls by 55.1% and token usage by 45.3% relative to DEER. Code, configurations, and reproducibility artifacts are available at `https://anonymous.4open.science/r/charcon-ner`.

1 INTRODUCTION

Named entity recognition (NER) identifies entities in unstructured text and assigns each entity an appropriate type, providing structured information for downstream applications. Large language models (LLMs) adapt extraction behavior in in-context learning (ICL) through instructions and labeled demonstrations, without task-specific gradient updates or model changes (Brown et al., 2020; Chen et al., 2023). This setting is particularly relevant to legal information extraction, where annotations are costly and entity schemas are domain-specific. Chinese legal NER, however, remains challenging because exact entity boundaries and context-dependent types must be determined together. Implicit word boundaries make small boundary deviations invalid, while similar monetary expressions or objects may correspond to different event-specific roles (Cao et al., 2022). Reliable in-context learning (ICL) therefore requires evidence for both exact entity boundaries and context-dependent entity types.

These requirements expose a mismatch between the evidence commonly supplied by ICL pipelines and the evidence needed to identify the correct entities and assign their types. At retrieval, whole-sequence similarity can emphasize topical relevance, while word-level lexical matching depends on segmentation units that may obscure short boundary- or role-sensitive constructions (Liu et al., 2022; Robertson & Zaragoza, 2009). At output, many methods generate extracted entity strings and their types as a separate list. Because this representation no longer preserves the original sentence, each predicted entity must later be matched back to its occurrence in the source text. At typing, independent label descriptions characterize categories individually but do not directly expose contextual distinctions among confusable legal roles. Recent systems strengthen different parts of the inference process through learned retrieval, targeted refinement, or multi-stage reasoning (Zhang et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2024; Bai et al., 2025; Wang et al., 2025; Mu et al., 2026). These observations motivate the question of how demonstration selection, source-preserving prediction, and role-sensitive evidence can be coordinated within an ICL framework with a frozen LLM and no task-specific gradient updates.

We view these limitations as an *evidence-alignment* problem in in-context NER. We define evidence alignment through three linked requirements. Retrieved evidence should capture local information relevant to entity boundaries. Source-context evidence should remain connected to the exact occurrence being predicted. Type-discriminative evidence should distinguish among plausible competing labels for the current decision.

We take a step towards solving this problem by proposing **C^3-NER**, a training-free framework that coordinates three evidence interfaces, namely Character-Pattern Retrieval, Copy-Preserving Annotation, and Contrastive Guidance. Character-Pattern Retrieval captures local constructions without relying on fixed word segmentation. Copy-Preserving Annotation keeps entity spans anchored to their source occurrences throughout extraction. Contrastive Guidance provides query-matched contextual evidence for distinguishing confusable entity types. Our main contributions are as follows. First, we introduce evidence alignment as a design principle for training-free in-context NER, and formalize the three interfaces that connect demonstration selection, source-preserving prediction, and role-sensitive type disambiguation. Second, we instantiate this principle in C^3-NER through overlapping character-pattern retrieval, source-grounded copy-preserving annotation, and query-activated contrastive guidance, followed by deterministic validation and span recovery. Third, we conduct extensive experiments on four NER benchmarks, including controlled retrieval, component ablation, matched-backbone, and inference-cost analyses. Extensive experiments across four NER benchmarks compare C^3-NER with five reproduced end-to-end systems. On the Chinese legal NER benchmark CAIL2021, C^3-NER reaches 90.99 F1, exceeding the strongest reproduced baseline DEER by 3.65 F1 points. On MSRA, it reaches 91.56 F1, improving over DEER by 1.87 F1 points. These gains come with a single LLM call per instance: on CAIL2021, C^3-NER uses 525 calls and 1,077,278 tokens, reducing LLM calls by 55.1% and token usage by 45.3% relative to DEER. Controlled retrieval and component ablation experiments evaluate the contributions of the three coordinated designs, while matched-backbone and inference-cost analyses characterize the scope and resource use of C^3-NER.

2 Related Work

Related work relevant to our study spans three complementary directions, generative and in-context NER, demonstration selection for ICL, and Chinese legal NER. Together, these directions characterize how entity predictions are represented, what labeled evidence is exposed during inference, and what boundary- and role-sensitive evidence is required by the target task.

Generative and In-Context NER Generative information extraction formulates structured prediction as sequence generation, with representative approaches including unified sequence generation (Yan et al., 2021), augmented natural language (Paolini et al., 2021), and schema-conditioned generation (Lu et al., 2022b). LLM-based NER further adapts extraction through demonstrations, entity markers, multi-turn interaction, or self-annotated examples (Lee et al., 2022; Chen et al., 2023; Shen et al., 2023; Wei et al., 2023; Xie et al., 2024). Recent systems enrich this process with consistency evidence, learned retrieval, refinement, or agent interaction (Wu et al., 2024; Zhang et al., 2024; Bai et al., 2025; Wang et al., 2025; Mu et al., 2026). For example, DEER combines label-guided demonstration evidence with targeted refinement over error-prone predictions (Bai et al., 2025). These studies show that both output representation and inference-time evidence influence NER performance. In contrast, C^3-NER directly targets this requirement through Copy-Preserving Annotation, which serializes entities over their original source sequences and thereby preserves a reliable correspondence between predicted spans and their source occurrences throughout inference.

Demonstration Selection for ICL Demonstration selection determines which labeled evidence is available to an in-context learner. Prior analyses further show that demonstrations convey information beyond direct input–label mappings, including the input distribution, label space, output format, and ordering (Min et al., 2022; Lu et al., 2022a). This suggests that demonstration relevance depends on what evidence the downstream decision requires. KATE ranks examples through semantic similarity (Liu et al., 2022), while lexical BM25 (Robertson & Zaragoza, 2009) and model-

aware retrievers such as EPR, Dr.ICL, and MDR (Rubin et al., 2022; Luo et al., 2024; Wang et al., 2024) use different representations and relevance signals. Other work incorporates diversity, iterative refinement, or topic coverage (Yang et al., 2023; Qin et al., 2024; Kweon et al., 2025), and retrieval-enhanced information extraction constructs query-dependent prompts from labeled examples (Shiraishi et al., 2024; Ge et al., 2026; Chen et al., 2025). These approaches improve which examples are selected, but retrieval relevance also depends on how the examples themselves are represented. To avoid this granularity mismatch, C^3-NER uses Character-Pattern Retrieval, which selects demonstrations through overlapping character patterns and thus retains boundary-sensitive local evidence without committing to fixed word segmentation.

Chinese Legal NER Chinese NER must identify entity boundaries without explicit word segmentation, motivating character- and lexicon-aware approaches such as Lattice LSTM and FLAT (Zhang & Yang, 2018; Li et al., 2020). Legal NER introduces an additional challenge because many entity types depend on their roles within legal events and case descriptions (Zhong et al., 2020; Chalkidis et al., 2022; Premasiri et al., 2025). In CAILIE, participants, objects, and monetary expressions can have similar surface forms while receiving different entity types according to their contextual functions (Cao et al., 2022). Chinese legal NER therefore couples precise boundary identification with context-dependent type discrimination. C^3-NER therefore aligns these coupled requirements within a single training-free framework. In this framework, character-pattern retrieval supplies locally relevant demonstrations, copy-preserving annotation preserves source occurrence, and query-activated contrastive guidance exposes role-sensitive distinctions among confusable legal types.

Taken together, prior work leaves three closely related challenges for in-context Chinese legal NER. Demonstrations must be retrieved at a granularity that preserves evidence for precise entity boundaries, predicted entities must remain connected to their occurrences in the source text, and context-dependent entity types require evidence that distinguishes confusable roles. These challenges are typically addressed at different stages of the prediction process. We study how the corresponding evidence can instead be coordinated around the same entity decision, motivating the evidence-alignment perspective developed in C^3-NER.

3 METHOD

Task setting We focus on few-shot Chinese legal NER, where the goal is to identify exact entity spans and assign their legal entity types from a Chinese legal text. Given a source-character sequence $x = (c_1, \ldots, c_L)$ and a legal entity type inventory $\mathcal{T}$, the task is to predict a set of typed spans $Y = \{(s_j, e_j, t_j)\}_{j=1}^{m}$, where $[s_j, e_j)$ denotes the entity boundaries and $t_j \in \mathcal{T}$ denotes its type. Exact-match evaluation requires both the predicted span boundaries and entity type to match the gold annotation, and the evaluated gold entities are non-overlapping. A labeled training set $\mathcal{D} = \{(x_i, Y_i)\}_{i=1}^{N}$ provides demonstrations and training-derived evidence for a query x_q. Neither the retriever nor the frozen LLM f_θ receives gradient-based updates, and development or test annotations are unavailable at prediction time.

3.1 OVERVIEW

We organize C^3-NER around *evidence alignment* with three coordinated components, **Character-Pattern Retrieval**, **Copy-Preserving Annotation**, and **Contrastive Guidance**. Figure 1 summarizes the workflow, in which demonstration evidence and role evidence are constructed separately and combined at inference.

Character-Pattern Retrieval. Given the query and the labeled demonstration pool, this stage retrieves an ordered set of examples whose local character-level constructions are most relevant to the current input.

Copy-Preserving Annotation. The retrieved examples are then serialized directly over their original source sequences, producing demonstration pairs that preserve entity occurrence, span boundaries, and surrounding context.

Contrastive Guidance. In parallel, a reusable Pattern Bank is constructed from labeled training contexts of predefined confusable types. For each query, Query-Level Activation selects matched patterns from this bank to form query-specific evidence for type discrimination.

At inference time, the source-aligned demonstrations and activated contrastive evidence are assembled with the query into a fixed prompt and passed to the frozen LLM. Deterministic validation and span recovery then map the model output back to typed source spans.

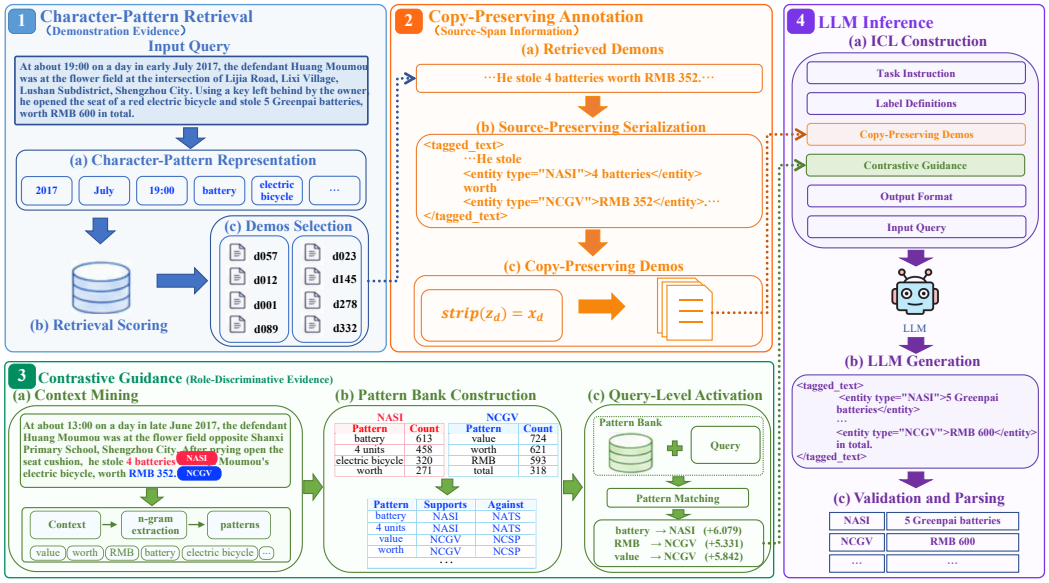

Figure 1: Overview of the evidence-aligned C^3-NER framework.

3.2 Character-Pattern Retrieval

For exact-span Chinese NER, useful demonstrations may be determined by short lexical and boundary-sensitive constructions rather than only sentence-level semantic similarity. Word-level lexical retrieval introduces an additional segmentation decision whose units need not coincide with annotated entity boundaries. We therefore represent both queries and candidate demonstrations with overlapping character patterns. For selected orders $\mathcal{N}$, let

$$G_n(x) = [c_i \cdots c_{i+n-1} \mid 1 \le i \le L - n + 1] \tag{1}$$

be the multiset of contiguous character patterns, and define

$$G_{\mathcal{N}}(x) = \biguplus_{n \in \mathcal{N}} G_n(x), \qquad U_{\mathcal{N}}(x) = g \mid g \in G_{\mathcal{N}}(x), \tag{2}$$

where $\biguplus$ denotes multiset union and $U_{\mathcal{N}}$ contains distinct patterns. Lower-order patterns retain characters that may participate in entity boundaries, whereas higher-order patterns capture short lexical constructions, numerical expressions, modifiers, and local contextual fragments. The representation therefore preserves local lexical evidence without introducing fixed word segmentation.

Given this representation, retrieval uses BM25-style lexical scoring (Robertson & Zaragoza, 2009),

$$S(x_q, d) = \sum_{g \in U_{\mathcal{N}}(x_q)} \mathrm{IDF}(g) \frac{\mathrm{tf}(g,d)(k_1+1)}{\mathrm{tf}(g,d) + k_1 \left(1 - b + b \frac{|G_{\mathcal{N}}(d)|}{\mathrm{avgdl}}\right)}. \tag{3}$$

Here, tf counts candidate-side occurrences, avgdl is mean pattern-document length, and $\mathrm{IDF}(g) = \log(1 + (N - \mathrm{df}(g) + 0.5)/(\mathrm{df}(g) + 0.5))$. The parameters k_1 and b control saturation and length normalization, while summing over distinct query patterns avoids repeated query-side contributions. We retain the top-K demonstrations in decreasing score order, break ties by training-set index, and fill any shortage of positive-score examples with zero-score examples in source order. The output of this stage is therefore an ordered demonstration set whose local character evidence is relevant to x_q; the selected examples are passed unchanged in this order to the copy-preserving representation stage.

3.3 COPY-PRESERVING ANNOTATION

The retrieved demonstrations still require an output representation that preserves which source occurrence is annotated. Mention-centric outputs instead generate detached strings or records, so occurrence localization must be reconstructed when a mention repeats or its boundary is ambiguous. We therefore serialize each retrieved (x_d, Y_d) directly over its original source sequence. Specifically, we construct $z_d = \text{Serialize}(x_d, Y_d)$ by inserting `<entity type="TYPE">mention</entity>` at the annotated spans and enclosing the sequence in a `<tagged_text>` wrapper.

Serialization follows source positions and satisfies the invariant

$$\text{strip}(z_d) = x_d, \tag{4}$$

where stripping removes structural tags only. This constraint preserves each entity's source occurrence, annotation extent, and surrounding context rather than requiring a separate mention-to-source alignment step. The same schema is required when generating the query output, so valid predictions can later be mapped directly back to typed source spans. The output of this stage is an ordered set of source-aligned demonstration pairs (x_d, z_d), which forms the demonstration portion of the final prompt.

3.4 CONTRASTIVE GUIDANCE

Even with locally relevant and source-aligned demonstrations, surface form alone may be insufficient when legal entity types are distinguished by different semantic or event roles. We therefore construct a reusable Pattern Bank that captures contextual evidence distinguishing predefined confusable types. Confusion groups $\mathcal{C} \subseteq \mathcal{T}$ are defined from the dataset schema before evaluation and remain fixed. For each entity in a group, we collect up to h characters on each side and extract patterns of orders $\mathcal{N}_c$. The mention itself is excluded, and left and right contexts are processed separately so that no pattern crosses the removed span.

Let $c_t(g)$ count pattern occurrences around type t, $N_t = \sum_g c_t(g)$, and $\mathcal{V}_\mathcal{C} = \bigcup_{t \in \mathcal{C}} g : c_t(g) > 0$. We compute

$$P(g \mid t) = \frac{c_t(g) + \alpha}{N_t + \alpha|\mathcal{V} * \mathcal{C}|}, \tag{5}$$

$$w(g; t_a, t_b) = \log \frac{P(g \mid t_a)}{P(g \mid t_b)}. \tag{6}$$

A positive $w(g; t_a, t_b)$ indicates that g occurs more strongly in the training contexts of t_a than of t_b. For each ordered pair $t_a \neq t_b$, we retain patterns with $c_{t_a}(g) \geq m$ and positive weight, rank them by weight and support, and keep at most M. Each Pattern Bank item (g, t_a, t_b, w) therefore records a contextual pattern, supported type, competing type, and contrastive strength. The resulting bank is constructed once from training data and remains fixed during evaluation.

Because only a subset of this corpus-level evidence is relevant to a particular query, the fixed bank is activated at query level rather than inserted wholesale into every prompt. For x_q, Query-Level Activation retains matched items according to

$$\mathcal{E}_\mathcal{C}(x_q) = \text{Top}_R\,(g, t_a, t_b, w) \in \mathcal{B}_\mathcal{C} \mid g \in \text{Substr}(x_q). \tag{7}$$

At most R items are retained per group, ordered by weight, then pattern length and support. Each item is rendered as a compact statement that g provides stronger training-context evidence for t_a than for t_b. Their union forms the query-specific guidance $\mathcal{G}(x_q)$. This stage therefore converts a reusable training-derived bank into role evidence matched to the current query, which is passed alongside the retrieved demonstrations to the final prompt.

3.5 LLM INFERENCE

The outputs of the preceding stages are assembled into a fixed prompt. With $\|$ denoting concatenation,

$$\begin{aligned} P(x_q) = I_{\text{task}} \,\|\, I_{\text{label}} \,\|\, [(x_{d_1} \,\|\, z_{d_1}) \,\|\, \cdots \,\|\, (x_{d_K} \,\|\, z_{d_K})] \\ \|\, \mathcal{G}(x_q) \,\|\, I_{\text{format}} \,\|\, x_q, \end{aligned} \tag{8}$$

Table 1: Statistics of the experimental datasets.

Dataset	Language	Domain	#Types	Train	Dev	Test	Eval.	Avg. Len.
CAIL2021	Chinese	Legal	10	4,197	525	525	525	64.13
MSRA	Chinese	News	3	46,364	–	4,365	200	46.18
CLUENER2020	Chinese	General	10	10,748	1,343	–	200	37.39
OntoNotes	English	General	18	59,924	8,528	8,262	200	18.11

where the demonstrations retain decreasing retrieval-score order, $\mathcal{G}(x_q)$ provides query-activated contrastive evidence, and I_{format} specifies the copy-preserving output schema. A single LLM call produces $z_q = f_\theta(P(x_q))$.

Before parsing, the response is required to contain well-formed tags, types from $\mathcal{T}$, and to satisfy $\text{strip}(z_q) = x_q$. Valid outputs are deterministically mapped to typed source spans using visible-character offsets; malformed, empty, unknown-type, or source-inconsistent predictions are rejected. If the copy condition fails, deterministic surface grounding retains only generated mentions that occur in the source.

4 Experiments

4.1 Datasets and Evaluation

CAIL2021 (Cao et al., 2022) is the primary benchmark. Because its original test partition lacks gold annotations, the labeled data is repartitioned 8:1:1 using a fixed random seed, yielding 4,197/525/525 training/development/test instances without identical-context duplication across partitions. Only the K and pattern-order sensitivity studies use development data; all other main experiments use test set. Fixed 200-instance subsets of **MSRA** (Levow, 2006), **CLUENER2020** (Xu et al., 2020), and **OntoNotes** (Weischedel et al., 2013) assess robustness across domains and languages. They are sampled once from the MSRA and OntoNotes test splits and the CLUENER2020 development split and reused across all conditions. Banks use the corresponding training split, and confusion groups are fixed before evaluation. We report entity-level micro precision, recall, and F1 under exact span-and-type matching.

4.2 Compared Methods

Complete-system comparisons reproduce **ConsistNER** (Wu et al., 2024), **EnDe** (Zhang et al., 2024), **CMAS** (Wang et al., 2025), **KDR-Agent** (Mu et al., 2026), and **DEER** (Bai et al., 2025), while retaining their characteristic workflows. All methods use the same partitions, schemas, evaluator, and DeepSeek V4 Flash backbone at every generative stage. Among these systems, DEER provides the closest comparison to our setting because it combines label-guided demonstration retrieval with targeted error reflection in a ICL workflow (Bai et al., 2025).

Controlled retrieval compares **Random**, **KATE** (Liu et al., 2022), **Word BM25** (Robertson & Zaragoza, 2009), and **Character-Pattern Retrieval** while fixing the demonstration budget, output format, Pattern Bank, Query-Level Activation, LLM, decoding, and parser. KATE uses `paraphrase-multilingual-MiniLM-L12-v2`; Word BM25 uses Jieba precise-mode segmentation for Chinese and whitespace tokenization for English. Word BM25 and Character-Pattern Retrieval share the BM25 scoring principle but differ in retrieval representation. Backbone comparisons change only the LLM.

4.3 Implementation Details

Generation uses temperature 0 and at most 512 output tokens, with $K = 8$. Character-Pattern Retrieval uses $\mathcal{N} = \{1, 2, 3\}$, $k_1 = 1.5$, and $b = 0.75$; Word BM25 shares the latter two settings. Guidance uses $h = 8$, $\mathcal{N}_c = \{1, 2, 3\}$, $\alpha = 1$, $m = 2$, $M = 10$, and $R = 3$. Settings remain fixed across datasets except schema-dependent confusion groups. GLM-5.2 and Qwen3.8-Flash use the same decoding settings.

5 RESULTS

The evidence-alignment view leads to three empirical expectations. A retrieval representation matched to local span decisions should improve demonstration selection under a fixed downstream interface, source-preserving generation should improve prediction while maintaining occurrence correspondence, and role-specific guidance should provide additional value after relevant demonstrations have been selected. We first establish the utility of the complete workflow and then test these expectations through controlled retrieval and component-level analyses.

5.1 EFFECTIVENESS AND CONTROLLED RETRIEVAL

Table 2 addresses two related but distinct questions by evaluating the effectiveness of the complete C^3-NER workflow under the unified protocol and the contribution of its retrieval design when the downstream interface is held fixed.

Table 2: End-to-end and controlled retrieval results under strict entity-level micro P/R/F1. Block A compares complete workflows; Block B varies only demonstration selection under a fixed downstream interface.

	CAIL2021			MSRA			CLUENER2020			OntoNotes		
Method	P	R	F1	P	R	F1	P	R	F1	P	R	F1
A. Complete end-to-end systems												
CMAS	64.29	55.80	59.75	67.61	52.96	59.40	48.76	50.11	49.42	62.45	49.36	55.14
EnDe	74.57	82.32	78.26	62.38	71.85	66.78	58.24	74.31	65.30	59.17	73.87	65.71
KDR-Agent	76.89	84.56	80.54	65.63	54.44	59.51	55.89	63.48	59.44	69.49	66.13	67.77
ConsistNER	78.60	83.19	80.83	71.67	69.82	70.73	63.58	70.06	66.67	70.93	71.61	71.27
DEER	86.27	88.43	87.34	89.19	**90.19**	89.69	72.87	**79.83**	**76.19**	**85.44**	**87.10**	**86.26**
C^3-NER (Ours)	**90.60**	**91.39**	**90.99**	**93.79**	89.44	**91.56**	**74.42**	74.74	74.58	82.90	81.29	82.09
B. Controlled demonstration retrieval												
Random	81.88	78.68	80.25	72.55	70.00	71.25	62.50	63.69	63.09	77.50	70.00	73.56
KATE	89.40	88.58	88.99	87.31	68.82	76.97	65.08	63.69	64.38	81.08	77.42	79.21
Word BM25	87.92	88.35	88.14	86.28	88.52	87.39	72.55	**76.86**	**74.64**	**83.06**	80.65	81.83
C^3-NER (Ours)	**90.60**	**91.39**	**90.99**	**93.79**	**89.44**	**91.56**	**74.42**	74.74	74.58	82.90	**81.29**	**82.09**

Overall effectiveness. Across the complete workflows in Block A, C^3-NER achieves the highest F1 on CAIL2021 and MSRA and the second-highest on CLUENER2020 and OntoNotes. On the primary CAIL2021 benchmark, C^3-NER reaches 90.99 F1 and exceeds DEER, the strongest reproduced baseline, by 3.65 points, while on MSRA it reaches 91.56 F1 and improves over DEER by 1.87 points. The consistently strong performance of DEER and C^3-NER across the four datasets highlights a shared advantage of making task-specific, query-dependent evidence construction an explicit part of inference, which gives the model localized and label-relevant evidence for entity decisions. Despite this shared emphasis, the two systems intervene at different stages of inference. DEER uses label-derived token statistics to prioritize entity-relevant evidence and subsequently revisits error-prone tokens through targeted reflection (Bai et al., 2025), whereas C^3-NER organizes character-level retrieval evidence, source occurrence, and role contrasts before generation so that the span-and-type decision is made within a single evidence-aligned inference step. The stronger results on CAIL2021 and MSRA further indicate that this pre-generation organization is particularly effective when exact entity decisions depend on local character constructions, while the role-sensitive structure of CAIL2021 additionally benefits from explicit source and type evidence.

Controlled retrieval. Block B directly evaluates the evidence-selection component by varying only demonstration retrieval while keeping the demonstration budget, output format, Pattern Bank, Query-Level Activation, LLM, decoding, and parser fixed. In this block, C^3-NER uses Character-Pattern Retrieval as its selector. Under the controlled interface, C^3-NER outperforms sentence-level KATE on all four datasets and achieves the highest F1 on CAIL2021, MSRA, and OntoNotes. Word BM25 provides the most direct comparison of retrieval granularity because both selectors

use the same lexical relevance principle while representing the input at different units. Replacing segmented words with overlapping character patterns improves F1 by 2.86 points on CAIL2021 and 4.17 points on MSRA, while maintaining closely matched performance on CLUENER2020 and OntoNotes. These results support segmentation-independent character granularity as an effective retrieval representation for preserving local evidence relevant to exact span-and-type prediction.

5.2 Component Attribution

Table 3 evaluates the three evidence interfaces on the CAIL2021 test split and measures copy preservation during generation. In Panel A, Character-Pattern Retrieval is replaced by Word BM25, Copy-Preserving Annotation by mention-centric JSON, and Contrastive Guidance is removed, with all remaining settings fixed.

Table 3: Component attribution and copy-preservation evidence.

A. Component ablation

Variant	P	R	F1
C^3-NER	**90.60**	91.39	**90.99**
w/o Character-Pattern Ret.	87.92	88.35	88.14
w/o Copy-Preserving Ann.	88.48	**91.50**	89.96
w/o Contrastive Guidance	89.61	91.24	90.42

Ret. = Retrieval; *Ann.* = Annotation;

B. Copy preservation

Dataset	Valid/Eval.	Rate
CAIL2021	517/525	98.48%
MSRA	189/200	94.50%
CLUENER2020	196/200	98.00%
OntoNotes	198/200	99.00%

Panel A reveals distinct contributions from the three evidence interfaces. Replacing Character-Pattern Retrieval with Word BM25 decreases F1 by 2.86 points even though both use the same BM25-style lexical scoring, showing that overlapping character patterns improve retrieval by preserving boundary-sensitive and local evidence beyond fixed word segments. Replacing Copy-Preserving Annotation with mention-centric JSON separates generated entities from the source sequence, while our representation inserts annotations directly into the original text and decodes them back to source spans; the resulting precision difference indicates that preserving this source connection improves localization decisions. Removing Contrastive Guidance eliminates query-matched evidence between competing roles and mainly reduces precision, showing its contribution to finer type discrimination. Together, retrieval determines which local evidence is exposed, copy preservation keeps predictions tied to their source spans, and contrastive guidance resolves the final role distinction.

Beyond its predictive contribution in Panel A, Copy-Preserving Annotation introduces an explicit source-correspondence property, which **Panel B** evaluates directly. Direct copy validity ranges from 94.50% to 99.00% across the four datasets, including 517 valid outputs among 525 CAIL2021 test instances. The consistently high validity shows that the source-preserving interface reliably maintains the original sequence during generation, complementing the F1 improvement observed when Copy-Preserving Annotation is retained. Together with the controlled retrieval comparison, these results connect the three interfaces of C^3-NER to the evidence-alignment principle by showing that Character-Pattern Retrieval selects locally relevant evidence, Copy-Preserving Annotation preserves source occurrence, and Contrastive Guidance supplies query-matched evidence for confusable roles.

5.3 Backbone Robustness

Table 4: Matched-backbone F1 on CAIL2021. ΔF1 denotes the improvement of C^3-NER over DEER.

Backbone	DEER	**C^3-NER**	ΔF1
DeepSeek V4 Flash	87.34	**90.99**	+3.65
GLM-5.2	87.06	**90.88**	+3.82
Qwen3.8-Flash	87.39	**90.81**	+3.42

Table 4 evaluates whether the effectiveness of the complete C^3-NER workflow is maintained when the underlying LLM changes. The DEER and C^3-NER workflows are retained in full, with only the backbone varied across **DeepSeek V4 Flash**, **GLM-5.2**, and **Qwen3.8-Flash**. Under these matched settings, C^3-NER consistently improves over DEER by 3.42–3.82 F1 points, while its absolute F1 remains stable between 90.81 and 90.99. The consistent gains across all three evaluated backbones show that the effectiveness of the complete evidence-aligned workflow is maintained across different generators, strengthening the system-level results observed with the primary backbone.

5.4 Inference Cost

Having established effectiveness across the evaluated backbones, we next examine whether these gains require additional generative interaction. We measure inference cost using the number of LLM calls and total input–output tokens over each evaluation set. Table 5 compares the complete workflows under the same evaluation settings.

Table 5: LLM interaction cost of complete workflows.

	CAIL2021		MSRA		CLUENER2020		OntoNotes	
Method	Calls	Tokens	Calls	Tokens	Calls	Tokens	Calls	Tokens
CMAS	4,600	7,232,720	3,300	1,747,621	3,300	1,880,027	3,300	2,423,111
EnDe	525	2,263,635	200	444,363	200	445,408	200	514,142
KDR-Agent	2,071	2,463,488	790	1,035,781	790	583,676	738	773,968
ConsistNER	5,247	2,800,501	4,597	879,530	4,597	1,402,651	4,597	2,476,665
DEER	1,170	1,970,484	345	437,922	351	311,524	252	365,203
C^3-NER	**525**	**1,077,278**	**200**	**233,343**	**200**	**214,523**	**200**	**297,356**

Across all four datasets, C^3-NER uses one LLM generation per instance and achieves the lowest token usage among the evaluated complete workflows. On CAIL2021, this corresponds to 55.10% fewer LLM calls and 45.30% fewer tokens than DEER. The difference follows directly from where each workflow places its inference operations. CMAS distributes extraction across multiple agents, KDR-Agent introduces reflective revision, and DEER performs targeted refinement after the initial extraction, so part of their decision process is carried out through additional generative interactions. C^3-NER instead constructs demonstration and role evidence before generation and handles validation, span recovery, and fallback grounding deterministically after generation, leaving the LLM to perform a single span-and-type decision for each query. EnDe also uses one generation per instance, yet C^3-NER consumes fewer tokens across all four datasets, showing that interaction depth alone does not determine inference cost and that the amount of information processed within each generation also matters. The resulting efficiency is therefore a structural consequence of concentrating evidence construction before generation and keeping the generative stage focused on the final entity decision.

6 Conclusion

This paper presented C^3-NER, a training-free framework for Chinese legal NER built around the principle of Evidence Alignment. Rather than treating demonstration selection, span generation, and type discrimination as independent stages, C^3-NER coordinates Character-Pattern Retrieval, Copy-Preserving Annotation, and Contrastive Guidance so that locally relevant demonstrations, source-span information, and role-discriminative evidence support the same exact span-and-type decision. Across four NER benchmarks, C^3-NER achieves strong end-to-end performance, while controlled retrieval, component ablation, matched-backbone, and inference-cost analyses support the effectiveness of the evidence-alignment design. The current study is limited in small external subsets, manually defined confusion groups, and the lack of multilingual legal NER evaluation. Future work will explore adaptive confusion-group construction, type-aware retrieval, cross-lingual legal NER, and privacy and bias assessment in legal-domain extraction.

AI USE STATEMENT

Generative AI tools were used solely for language editing and polishing to improve the readability and clarity of the manuscript. We did not use generative AI tools for any research tasks requiring mandatory disclosure under the ICLR AI Policy for Authors, including methodology or experiment design, method implementation, data processing or analysis, result interpretation, hypothesis formulation, or mathematical reasoning. All AI-assisted edits were reviewed by the authors. The authors take full responsibility for the final content of this work, including all text, claims, results, and artifacts.

ETHICS STATEMENT

This work uses existing benchmark datasets for methodological evaluation and does not involve new data collection or human-subject experimentation. The proposed framework is studied as an information extraction method and is not intended for autonomous legal decision-making. Any deployment in legal or other high-stakes settings would require domain-specific validation and appropriate human oversight.

REFERENCES

Fan Bai, Hamid Hassanzadeh, Ardavan Saeedi, and Mark Dredze. Llms are better than you think: Label-guided in-context learning for named entity recognition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025*, pp. 28372–28392. ACL, 2025.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*. OpenReview.net, 2020.

Yu Cao, Yuanyuan Sun, Ce Xu, Chunnan Li, Jinming Du, and Hongfei Lin. CAILIE 1.0: A dataset for challenge of AI in law - information extraction V1.0. *AI Open*, 3:208–212, 2022.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL 2022*, pp. 4310–4330. ACL, 2022.

Jiawei Chen, Yaojie Lu, Hongyu Lin, Jie Lou, Wei Jia, Dai Dai, Hua Wu, Boxi Cao, Xianpei Han, and Le Sun. Learning in-context learning for named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, ACL 2023*, pp. 13661–13675. ACL, 2023.

Zihan Chen, Lei Shi, Weize Wu, Qiji Zhou, and Yue Zhang. 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, EMNLP 2025*, pp. 25162–25176. ACL, 2025.

Yao Ge, Yuting Guo, Sudeshna Das, and Abeed Sarker. Improving few-shot named entity recognition for large language models using structured dynamic prompting with retrieval augmented generation. *npj Artificial Intelligence*, 2(39), 2026.

Wonbin Kweon, SeongKu Kang, Runchu Tian, Pengcheng Jiang, Jiawei Han, and Hwanjo Yu. Topic coverage-based demonstration retrieval for in-context learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025*, pp. 19900–19912. ACL, 2025.

Dong-Ho Lee, Akshen Kadakia, Kangmin Tan, Mahak Agarwal, Xinyu Feng, Takashi Shibuya, Ryosuke Mitani, Toshiyuki Sekiya, Jay Pujara, and Xiang Ren. Good examples make a faster learner: Simple demonstration-based learning for low-resource NER. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL 2022*, pp. 2687–2700. ACL, 2022.

Gina-Anne Levow. The third international chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proceedings of the Fifth Workshop on Chinese Language Processing, SIGHAN@COLING/ACL 2006*, pp. 108–117. ACL, 2006.

Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuanjing Huang. FLAT: Chinese NER using flat-lattice transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pp. 6836–6842. ACL, 2020.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? In *The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, DeeLIO@ACL 2022*, pp. 100–114. ACL, 2022.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL 2022*, pp. 8086–8098. ACL, 2022a.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. Unified structure generation for universal information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL 2022*, pp. 5755–5772. ACL, 2022b.

Man Luo, Xin Xu, Zhuyun Dai, Panupong Pasupat, Mehran Kazemi, Chitta Baral, Vaiva Imbrasaite, and Vincent Y. Zhao. Dr.ICL: Demonstration-retrieved in-context learning. *Data Intell.*, 6(4): 909–922, 2024.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*, pp. 11048–11064. ACL, 2022.

Wenxuan Mu, Jinzhong Ning, Di Zhao, and Yijia Zhang. A multi-agent LLM framework for multi-domain low-resource in-context NER via knowledge retrieval, disambiguation and reflective analysis. In *Fortieth AAAI Conference on Artificial Intelligence, AAAI 2026*, pp. 32528–32536. AAAI, 2026.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. Structured prediction as translation between augmented natural languages. In *International Conference on Learning Representations, ICLR 2021*. OpenReview.net, 2021.

Damith Premasiri, Tharindu Ranasinghe, Ruslan Mitkov, Mo El-Haj, and Ingo Frommholz. Survey on legal information extraction: current status and open challenges. *Knowl. Inf. Syst.*, 67(12): 11287–11358, 2025.

Chengwei Qin, Aston Zhang, Chen Chen, Anirudh Dagar, and Wenming Ye. In-context learning with iterative demonstration selection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7441–7455. ACL, 2024.

Stephen E. Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, 2009.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022*, pp. 2655–2671. ACL, 2022.

Yongliang Shen, Zeqi Tan, Shuhui Wu, Wenqi Zhang, Rongsheng Zhang, Yadong Xi, Weiming Lu, and Yueting Zhuang. Promptner: Prompt locating and typing for named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, ACL 2023*, pp. 12492–12507. ACL, 2023.

Enzo Shiraishi, Raphael Y. de Camargo, Henrique L. P. Silva, and Ronaldo C. Prati. Retrieval-enhanced named entity recognition. *CoRR*, abs/2410.13118, 2024.

Huazheng Wang, Jinming Wu, Haifeng Sun, Zixuan Xia, Daixuan Cheng, Jingyu Wang, Qi Qi, and Jianxin Liao. MDR: model-specific demonstration retrieval at inference time for in-context learning. In *Proceedings of the 2024 Conference of the North American Chapter of the ACL: Human Language Technologies, NAACL 2024*, pp. 4189–4204. ACL, 2024.

Zihan Wang, Ziqi Zhao, Yougang Lyu, Zhumin Chen, Maarten de Rijke, and Zhaochun Ren. A cooperative multi-agent framework for zero-shot named entity recognition. In *Proceedings of the ACM on Web Conference 2025, WWW 2025*, pp. 4183–4195. ACM, 2025.

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. Zero-shot information extraction via chatting with chatgpt. *CoRR*, abs/2302.10205, 2023.

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. OntoNotes release 5.0. LDC2013T19, 2013.

Chenxiao Wu, Wenjun Ke, Peng Wang, Zhizhao Luo, Guozheng Li, and Wanyi Chen. Consistner: Towards instructive NER demonstrations for llms with the consistency of ontology and context. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024*, pp. 19234–19242. AAAI, 2024.

Tingyu Xie, Qi Li, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 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 ACL: Human Language Technologies, NAACL 2024*, pp. 583–593. ACL, 2024.

Liang Xu, Yu Tong, Qianqian Dong, Yixuan Liao, Cong Yu, Yin Tian, Weitang Liu, Lu Li, and Xuanwei Zhang. CLUENER2020: fine-grained named entity recognition dataset and benchmark for chinese. *CoRR*, abs/2001.04351, 2020.

Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang, and Xipeng Qiu. A unified generative framework for various NER subtasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, ACL 2021*, pp. 5808–5822. ACL, 2021.

Zhao Yang, Yuanzhe Zhang, Dianbo Sui, Cao Liu, Jun Zhao, and Kang Liu. Representative demonstration selection for in-context learning with two-stage determinantal point process. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, pp. 5443–5456. ACL, 2023.

Meishan Zhang, Bin Wang, Hao Fei, and Min Zhang. In-context learning for few-shot nested named entity recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024*, pp. 10026–10030. IEEE, 2024.

Yue Zhang and Jie Yang. Chinese NER using lattice LSTM. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018*, pp. 1554–1564. ACL, 2018.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. How does NLP benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pp. 5218–5230. ACL, 2020.

A Additional Method Definitions

A.1 Retrieval and Recovery Details

Character-Pattern Retrieval ranks candidate demonstrations by decreasing retrieval score, with ties resolved by the original training-set index. If fewer than K candidates receive positive scores, zero-score examples are added in source order until the demonstration budget is reached.

For Copy-Preserving Annotation, each retrieved example is serialized with the wrapper `<tagged_text>`, and each entity is represented as `<entity type="TYPE">mention</entity>`. Tags are inserted in source-span order while preserving the surrounding source text and the retrieval order of demonstrations. At prediction time, outputs are checked for a valid wrapper, balanced entity tags, valid entity types, and exact reconstruction of the source sequence after tag removal. Empty mentions, unknown types, malformed tags, and source-inconsistent spans are rejected. If the copy condition is not satisfied, deterministic surface grounding retains only predicted mention strings that occur in the source text and discards ungroundable mentions.

A.2 Directional Evidence Storage

For each confusion group $\mathcal{C}$, patterns are counted separately on each side of the entity context while excluding the entity mention itself. A pattern associated with an ordered type pair (t_a, t_b) is eligible when its support for t_a satisfies the minimum-frequency requirement and its directional weight $w(g; t_a, t_b)$ is positive. The top M candidates according to weight and support form

$$\mathcal{B}_{\mathcal{C}} = \{(g, t_a, t_b, w)\}, \tag{9}$$

where each tuple stores a contextual pattern, its supported type, its competing type, and its directional weight. Query matches are ranked by weight, followed by pattern length and support, and at most R matched patterns are retained for each confusion group. The activated guidance is

$$\mathcal{G}(x_q) = \bigcup_{\mathcal{C}} \mathcal{E}_{\mathcal{C}}(x_q). \tag{10}$$

Each selected tuple is converted into a compact directional statement describing stronger training-context evidence for the supported type. Pattern Banks and confusion groups remain fixed throughout evaluation, and neither their construction nor query-level activation uses development or test annotations.

B Dataset and Reproduction Details

B.1 Datasets and Evaluation Protocol

CAIL2021 is repartitioned into training, development, and test sets with an 8:1:1 ratio while preventing identical contexts from crossing partitions. The 525-instance test split is used for complete-system comparison, controlled retrieval, component ablation, backbone evaluation, efficiency analysis, and diagnostics. The 525-instance development split is used exclusively for sensitivity analysis of the demonstration budget K and character-pattern orders.

The three external datasets use fixed 200-instance evaluation subsets that remain unchanged across methods and experimental conditions. Each dataset constructs its Pattern Bank only from its labeled training split, and the schema-defined confusion groups are fixed before evaluation. All reported NER results use exact-match micro precision, recall, and F1. A prediction is correct only when both its entity boundaries and type match the gold annotation.

B.2 Baseline Implementations

For controlled retrieval, Random selects demonstrations without query-dependent relevance scoring. KATE (Liu et al., 2022) embeds complete inputs with

Table 6: Effect of the number of demonstrations and corresponding token usage on the CAIL2021 development split.

K	P	R	F1	ΔF1	Tokens
0	59.984	57.549	58.742	-32.198	456,860
2	88.302	88.202	88.252	-2.688	614,511
4	89.468	89.909	89.688	-1.252	764,615
6	90.143	90.895	90.518	-0.422	922,394
8	**90.496**	91.388	**90.940**	0.000	1,077,242
10	90.319	91.313	90.813	-0.127	1,233,954
12	90.432	**91.426**	90.926	-0.014	1,396,414

Table 7: Effect of character-pattern orders on the CAIL2021 development split.

Orders	P	R	F1	ΔF1
$\{1\}$	90.440	91.161	90.799	-0.141
$\{1, 2\}$	90.392	91.009	90.699	-0.241
$\{1, 2, 3\}$	**90.496**	**91.388**	**90.940**	0.000
$\{1, 2, 3, 4\}$	90.034	91.161	90.594	-0.346

`paraphrase-multilingual-MiniLM-L12-v2` and ranks them using sentence-level semantic similarity. Word BM25 (Robertson & Zaragoza, 2009) uses Jieba precise-mode segmentation for Chinese and whitespace tokenization for English. Word BM25 and Character-Pattern Retrieval share $k_1 = 1.5$ and $b = 0.75$, with representation granularity providing the primary difference. The demonstration budget, Copy-Preserving Annotation, Pattern Bank, Query-Level Activation, backbone, decoding configuration, and deterministic parser remain fixed in the controlled comparison.

Complete-system baselines retain their characteristic workflows. ConsistNER (Wu et al., 2024) uses entity pre-identification, consistency-oriented demonstration selection, and extraction. EnDe (Zhang et al., 2024) uses a supervised multi-view retriever incorporating semantic, boundary, and label information. CMAS (Wang et al., 2025) coordinates multiple agents, KDR-Agent (Mu et al., 2026) combines knowledge-enhanced evidence with reflective revision, and DEER (Bai et al., 2025) uses label-guided evidence followed by targeted refinement. Public implementations are adapted to the target schemas and unified evaluator while preserving their retrieval, prompting, intermediate-generation, and refinement procedures. All main generative stages use DeepSeek V4 Flash. Matched-backbone comparisons retain each complete workflow and change only the underlying LLM.

C Hyperparameter Sensitivity

Tables 6 and 7 report sensitivity to the demonstration budget K and character-pattern orders. Both analyses use the CAIL2021 development split while keeping the remaining configuration unchanged.

Performance rises substantially once demonstrations are introduced and remains close across $K = 8$, 10, and 12, while token usage continues to increase. This places the fixed $K = 8$ setting within a stable high-performing region. For character-pattern orders, $\{1, 2, 3\}$ achieves the highest F1, while extending the representation to four-character patterns reduces performance by 0.346 F1 points.

D Inference Efficiency

Table 8 reports LLM calls and total token usage for the complete end-to-end systems across all four evaluation datasets. For C^3-NER, Pattern Bank construction, demonstration retrieval, Query-Level Activation, validation, parsing, and fallback do not require additional LLM calls.

Table 8: LLM calls and token usage of complete end-to-end workflows. Bold and underline denote the lowest and second-lowest token usage for each dataset, respectively.

Dataset	Method	Eval.	Calls	Tokens
CAIL2021	CMAS	525	4,600	7,232,720
	EnDe	525	525	2,263,635
	KDR-Agent	525	2,071	2,463,488
	ConsistNER	525	5,247	2,800,501
	DEER	525	1,170	<u>1,970,484</u>
	C^3-NER	525	525	**1,077,278**
MSRA	CMAS	200	3,300	1,747,621
	EnDe	200	200	444,363
	KDR-Agent	200	790	1,035,781
	ConsistNER	200	4,597	879,530
	DEER	200	345	<u>437,922</u>
	C^3-NER	200	200	**233,343**
CLUENER2020	CMAS	200	3,300	1,880,027
	EnDe	200	200	445,408
	KDR-Agent	200	790	583,676
	ConsistNER	200	4,597	1,402,651
	DEER	200	351	<u>311,524</u>
	C^3-NER	200	200	**214,523**
OntoNotes	CMAS	200	3,300	2,423,111
	EnDe	200	200	514,142
	KDR-Agent	200	738	773,968
	ConsistNER	200	4,597	2,476,665
	DEER	200	252	<u>365,203</u>
	C^3-NER	200	200	**297,356**

Table 9: Copy-preservation diagnostics across the four evaluation datasets.

Dataset	Eval.	Copy-valid	Valid rate
CAIL2021	525	517	98.48%
MSRA	200	189	94.50%
CLUENER2020	200	196	98.00%
OntoNotes	200	198	99.00%

On CAIL2021, C^3-NER uses 525 LLM calls and 1,077,278 tokens, corresponding to 55.1% fewer calls and 45.3% fewer tokens than DEER. The complete cross-dataset results provide the corresponding inference-cost comparison for the additional evaluation settings.

E Fine-Grained Diagnostics

E.1 Copy-Preservation Diagnostics

Copy validity directly evaluates whether Copy-Preserving Annotation satisfies the source-reconstruction constraint. A response is copy-valid when it passes the structural checks used by the deterministic parser and removing the annotation tags exactly reconstructs the source sequence.

Across the four datasets, 94.50%–99.00% of responses satisfy the source-reconstruction constraint directly. This diagnostic evaluates compliance with the representation-level invariant and does not by itself indicate whether the recovered entity boundaries or types are correct.

Table 10: Dataset-specific contrastive confusion groups used for corpus-induced evidence construction.

Dataset	Contrastive confusion groups
CAIL2021	{NATS, NASI}; {NCSP, NCSM, NCGV}; {NHCS, NHVI}; {NS, NO}.
MSRA	{PER, ORG}; {LOC, ORG}.
CLUENER2020	{company, government, organization}; {address, scene}; {book, movie, game}; {name, position}.
OntoNotes	{organization, geographical/social/political, location, facility}; {date, time}; {cardinal, ordinal, quantity, percent, money}; {work of art, product, event, law, language}; {person, national/religious/political}.

F CONTRASTIVE CONFUSION GROUPS

Confusion groups are defined from each dataset's schema before evaluation and specify the type families for which contrastive evidence is collected from labeled training contexts. They remain fixed throughout evaluation, and no label relations are adapted from development or test predictions.