

Structure Matters: Semantic-Structural Dual-Path Arbitration for Named Entity Recognition with LLMs

Xinghong Hou¹, Yiqi Tong^{2,*}, Wei Chen², Wei Guo², Fuzhen Zhuang^{2,*}

¹SWJTU-Leeds Joint School, Southwest Jiaotong University

²School of Artificial Intelligence, Beihang University

2023115876@my.swjtu.edu.cn, {yqtong, zhuangfuzhen}@buaa.edu.cn

Abstract

Named entity recognition (NER) extracts entity spans and assigns them to predefined types. Large language models (LLMs) are increasingly used for NER, but most LLM-based methods still rely on a mainly semantic view, judging whether a candidate span is compatible with an entity type. Although effective, this view can miss entities with weak type semantics, over-accept semantically salient modifiers, and leave little chance to recover spans omitted by early candidate extraction. We argue that entity spans are also traceable in sentence form: local structure and prepositions provide cues about a span’s position and function. We propose **SSDA-NER**, a **Semantic-Structural Dual-path Arbitration** framework that combines semantic understanding with structural analysis. Specifically, the Semantic path checks candidate type compatibility, while the Structure path withholds explicit semantic guidance during its first stage and identifies possible mentions through structural analysis. The Arbitration Agent then compares their rationales and confidence scores to adjudicate disagreements. Experiments on nine datasets with three backbone LLMs show that the two views are complementary: the Structure path substantially supplements entities missed by the Semantic path on some datasets, and arbitration outperforms existing multi-agent methods in most evaluated settings. Source code is available at <https://github.com/MrHoux/SSDA-NER>.

1 Introduction

Named entity recognition (NER) is a fundamental step for transforming unstructured text into structured knowledge and supporting downstream applications (Li et al., 2022a; Tong et al., 2021, 2026). Recent large language models (LLMs) have enabled a new line of prompt-driven NER methods, where LLMs are adapted to target schemas through

*Corresponding authors.

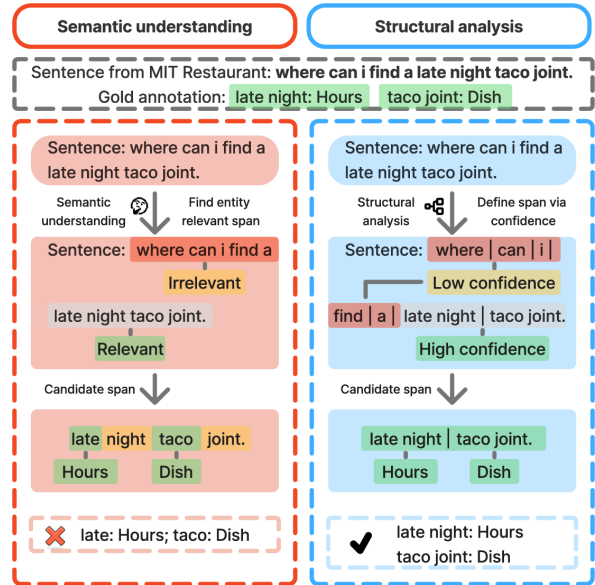


Figure 1: An example of disagreement between semantic understanding and structural analysis in NER.

instructions, in-context examples, and task reformulation (Liu et al., 2023; Gonzalez-Gallardo et al., 2023). These methods show that LLMs can exploit contextual information to improve entity typing and disambiguation, especially when labeled data are scarce (Fetahu et al., 2023; Tan et al., 2023). However, the context used in existing LLM-based NER methods remains largely semantic: it mainly helps the model judge whether a span is compatible with an entity type or whether a candidate prediction is plausible. Position and function information, although closely related to how entity mentions are formed in context, has received much less attention.

Existing LLM-based NER methods mainly follow two lines: task reformulation and semantic enhancement. Task reformulation methods, such as GPT-NER (Wang et al., 2025a) and ChatIE (Wei et al., 2024), convert NER into generation or question-answering formats (Tong et al., 2022). Decomposed-QA (Xie et al., 2023) further explores

syntactic prompting by decomposing sentences into smaller units, but its label-wise extraction still relies on semantic label matching rather than a separated structural analysis path. Semantic enhancement methods, such as CMAS (Wang et al., 2025b), KDR-Agent (Mu et al., 2026), and VerifiNER (Kim et al., 2024), improve extraction through contextual features, external knowledge, or verification. These methods differ in form and auxiliary information, but they still mainly use context semantically to judge whether a candidate span is plausible under an entity type.

In the NER task, semantic understanding and structural analysis do not merely duplicate each other; instead, they examine the same candidate span from different reasoning paths (Biales, 2023; Li et al., 2022b). The Semantic path may be conservative when the type-related meaning of a span is not explicit, whereas the Structure path may still identify it as a plausible entity mention according to its position and function in the sentence (Li et al., 2021; Shen et al., 2021). Conversely, the Semantic path may accept a span because it carries strong type-related meaning, while the Structure path may question whether the span is only a modifier or an incomplete part of a larger mention (Wu et al., 2022). Figure 1 illustrates how semantic understanding and structural analysis may lead to different span-level judgments. Such disagreements are not simply errors to be removed, but useful signals for multi-agent decision-making (Wang et al., 2023; Yao et al., 2023a; Du et al., 2024). When different agents are designed to produce comparable candidate spans through semantic understanding and structural analysis, their conflicts can expose uncertain cases that require further judgment (Yao et al., 2023b; Madaan et al., 2023). This also addresses a limitation of many existing NER pipelines, where candidate spans are produced from a single semantic view and later modules can only refine or verify what has already been proposed. By allowing the Semantic and Structure paths to disagree, the framework mitigates the risk of one-sided extraction and provides a more reliable basis for arbitration (Du et al., 2024).

To address this limitation, we propose **SSDA-NER**, a **Semantic-Structural Dual-path Arbitration** framework that combines semantic understanding with structural analysis. Instead of judging candidate spans only by semantic compatibility with the entity schema, SSDA-NER examines them from two complementary views: the Semantic path

checks whether the semantics of a span are appropriate, while the Structure path checks the span’s position and function in the sentence. Their disagreements are treated as localized uncertainty signals and handled by arbitration with rationales and confidence scores. In this way, SSDA-NER mitigates the one-sided limitation of semantic-only extraction.

Our contributions are threefold: (1) We define a structural path for LLM-based NER that withholds entity types and other semantic guidance, treating candidate strings as opaque units and focusing on structural analysis of their sentence position and function. (2) We propose a dual-path arbitration framework that separates semantic and structural judgments over candidate spans and uses their disagreements, rationales, and confidence scores to support localized decisions. (3) We provide experiments on nine NER benchmarks with three backbone LLMs, showing that semantic and structural analysis are complementary and that their arbitration improves over existing multi-agent NER baselines across most of the 27 evaluated settings.

2 Related Work

We review LLM-based NER methods that elicit entity predictions from general-purpose models through prompting and auxiliary reasoning. Existing studies mainly follow two lines: task reformulation, which adapts NER to LLM-friendly formats, and semantic enhancement, which strengthens semantic understanding during extraction.

Task Reformulation NER. Task reformulation methods redesign NER into formats that are easier for LLMs to follow. GPT-NER (Wang et al., 2025a) converts sequence labeling into text generation with special entity markers and self-verification. ChatIE (Wei et al., 2024) casts information extraction as a multi-turn question-answering process. CodeIE (Li et al., 2023) represents structured extraction outputs as code-like generation results. Decomposed-QA (Xie et al., 2023) decomposes NER into label-wise questions and examines prompting variants for extraction. Although effective, these methods still handle mention existence, span selection, and entity typing largely within a single semantic generation process. Our method keeps the prediction format secondary and introduces structural analysis as a parallel perspective.

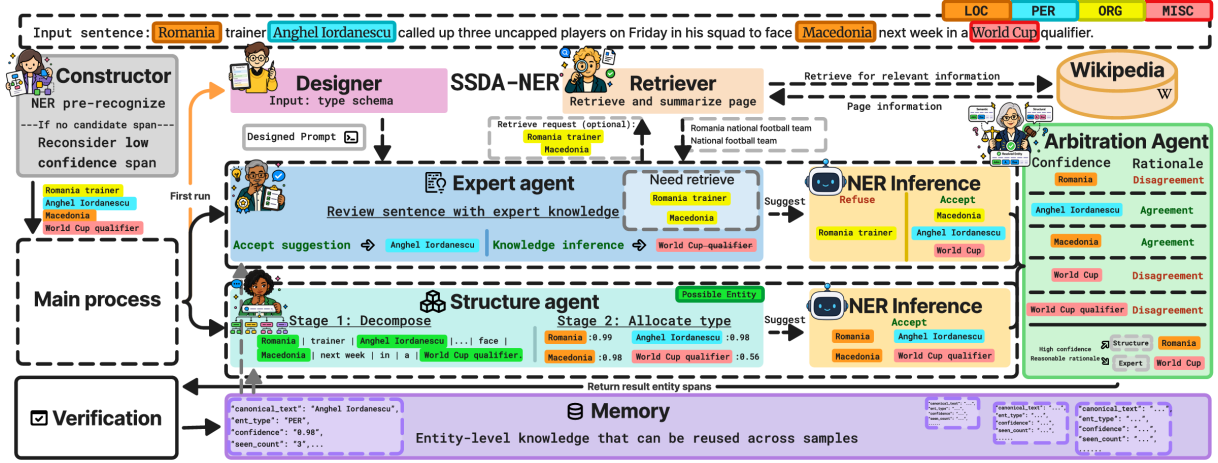


Figure 2: SSDA-NER pipeline on a CoNLL-2003 example.

Semantic Enhancement NER. Semantic enhancement methods improve LLM-based NER by providing richer semantic guidance. PromptNER (Ashok and Lipton, 2023) uses entity type definitions, explanations, and examples to improve candidate generation and type understanding. Self-Improving (Xie et al., 2024) constructs reliable self-annotated demonstrations from unlabeled corpora. DEER (Bai et al., 2025) retrieves demonstrations with label-guided token statistics and performs targeted correction. KDR-Agent (Mu et al., 2026) strengthens in-context NER through knowledge retrieval, entity disambiguation, and reflective analysis. CMAS (Wang et al., 2025b) uses multiple agents to model contextual type-related features and demonstration helpfulness. VerifNER (Kim et al., 2024) verifies and revises predictions through knowledge-grounded reasoning. These methods enrich semantic understanding, but still mainly judge whether a candidate span fits an entity type; the sentence’s structural analysis is rarely separated as an independent decision signal. Our framework separates semantic understanding from structural analysis and uses their disagreements to mitigate single-view extraction bias.

3 Methodology

3.1 Problem Formulation

Given an input sentence x and an entity schema $\mathcal{T}$, NER aims to produce a mention set

$$\mathcal{Y} = \{(s_i, e_i, t_i, \alpha_i)\}_{i=1}^N,$$

where (s_i, e_i) are character offsets in x , $t_i \in \mathcal{T}$ is the entity type, and $\alpha_i \in [0, 1]$ is a confidence score. We use the backbone LLM directly for each target dataset without target-specific fine-tuning.

3.2 Framework Overview

Figure 2 shows our Semantic-Structural Dual-Path Arbitration framework (SSDA-NER). A Constructor first builds candidate spans, while a Designer prepares schema-aware prompts. The Constructor uses generic low-confidence fallback spans only when the merged candidate set is empty. The Expert Agent performs semantic analysis and may ask the Retriever for Wikipedia information. In parallel, the Structure Agent first decomposes the sentence and then allocates entity types to possible spans. The two paths are converted by NER Inference into comparable candidate predictions with confidence scores. Agreed predictions can pass through directly, while disagreements and low-confidence spans are reconsidered by the Arbitration Agent. Verification checks the final format, schema, and offsets. Memory stores only deterministically verified entity-level records for later retrieval.

Agent handoffs are structured rather than free-form. Each intermediate record keeps the span id, exact offsets, candidate type, rationale, and a normalized self-reported confidence score. Confidence is used for routing and filtering, not as a calibrated probability.

3.3 Scope of Structural Analysis

SSDA-NER does not assume an external symbolic parser. In this paper, “structural” refers to LLM-guided analysis about a span’s local position and function after sentence decomposition: whether it is a referential mention, modifier, appositive unit, prepositional complement, or unit inside a larger expression.

The Structure path is not implemented as another semantic agent. Its first stage is a schema-withheld

structural pass: the model receives the sentence and candidate offsets, but not entity type names, label descriptions, examples and RAG facts. Content words and named strings are treated as opaque units indexed by position and offset, and the output schema only allows selected units, structural roles, structural cues, and `schema_used:false`. The retained signals are word order, offsets, punctuation, parenthetical or quoted regions, coordination/apposition markers, and prepositions. Only after this schema-free selection by position and function does the second stage reintroduce the entity schema for type allocation. This separation is therefore an observable interface constraint rather than a claim that the hidden states of the LLM contain no semantic information. Unlike generic multi-agent prompting or self-consistency, the two paths do not simply vote over the same semantic view.

3.4 Initial Candidate Span Construction

The Constructor creates an initial candidate set

$$\mathcal{C}_0 = \{c_j = (id_j, q_j, s_j, e_j)\}_{j=1}^M,$$

where $q_j = x[s_j : e_j]$. The Candidate Agent and direct-seed NER inference first propose spans independently. Their outputs are merged, aligned to exact offsets, and deduplicated. Only if this merged candidate set is empty does the Constructor add generic low-confidence fallback spans; if any candidate span already exists, no generic fallback span is introduced by the Constructor.

3.5 Dual-path Guidance

For each candidate c_j , both paths produce guidance

$$g_j^r = (B_j^r, T_j^r, R_j^r, \alpha_j^r), \quad r \in \{E, S\},$$

where B_j^r is a span decision, $T_j^r \subseteq \mathcal{T}$ is the type hint set, R_j^r is textual analysis or rationale, and α_j^r is confidence. Span proposals are added only when they are schema-compatible, quote-anchored, bounded in length, and above a confidence threshold. Structure Agent proposals use a stricter acceptance rule because they are more likely to introduce span variants.

Guidance is always attached to a candidate id or to a quote-anchored span proposal. This keeps later modules from comparing ungrounded text strings and allows Expert and Structure outputs to be aligned by id or offsets.

Semantic Guidance with Expert Agent. Before evaluation begins on a dataset, the Designer builds a dataset-level expert prompt. It serializes the entity schema, type descriptions, residual labels, and annotation constraints, then merges this information with the general Expert and Expert Retrieval templates. The resulting guidance is fixed for the dataset rather than generated per sentence. It tells the Expert how labels should be interpreted, what counts as an identity-bearing mention, which role nouns, descriptors, dates, or numerals should usually be rejected, and when residual labels apply only after specific labels are ruled out.

At inference time, the Expert Agent implements the Semantic path and performs a semantic review over candidate spans. For each span, it reads the sentence context, schema guidance, memory hints, and any RAG handoff, and returns candidate types, excluded types, direct textual support with offsets, a rationale, and a normalized self-reported confidence score. The decision is based on semantic compatibility between the span and the schema, local lexical and contextual cues, and explicit supporting or counter cues. Current-sentence support has priority: unsupported spans are assigned low confidence or abstained from, and any new span proposal must be anchored by an exact quote and legal type hints. When local context is insufficient to decide entity identity or type, an Expert Retrieval Planner may issue span-linked RAG requests. The Retriever queries Wikipedia for related content and the RAG handoff is treated as auxiliary support rather than as a final label.

Structure Guidance with Structure Agent. The Structure Agent is separated from the semantic Expert path and follows the two-stage process in Figure 2. It does not produce a symbolic parse tree, but it explicitly reasons over sentence-internal position and function under the schema-withheld interface defined above. *Stage 1: Decompose* is schema-free span-position selection: the agent sees the sentence, candidate offsets, and structure markers, but not entity type names, label descriptions, examples and RAG facts. It decomposes the sentence into local units using order, punctuation, parenthetical or quoted regions, coordination/apposition markers, and prepositional attachments. It then asks whether each opaque unit functions as a possible entity-bearing position, a modifier of another unit, part of a larger phrase, or unsupported text. This stage may keep an existing candidate, reject it as struc-

turally weak, or propose a more coherent possible entity span, but it only records structural analysis and never assigns an entity type in Stage 1.

Stage 2: Allocate type performs type allocation for the structurally selected spans. Type allocation is a matter of semantic interpretation in itself, and is therefore handled separately from Stage 1. Only after Stage 1 has identified possible mention positions does the Structure Agent compare those spans with the entity schema and assign candidate types and confidence. A span receives high confidence only when its structural analysis and schema compatibility agree; otherwise it is passed forward with low confidence. However, in order to retain the results of the structural analysis, results with low confidence will not be deleted and will be re-considered at NER inference. The final structural guidance therefore contains the selected span, its structural rationale, type hints, and confidence.

$$\mathcal{U}^S = D_\theta(x, \mathcal{C}; \mathcal{T} = \emptyset), \quad \mathcal{G}^S = A_\theta(x, \mathcal{U}^S; \mathcal{T}),$$

where D_θ denotes Stage 1 decomposition and A_θ denotes Stage 2 type allocation.

Path-level NER Inference. NER Inference is run separately with Expert guidance and Structure guidance, producing candidate prediction sets $\mathcal{Y}^E$ and $\mathcal{Y}^S$. Each candidate prediction has the same tuple format (s, e, t, α) , which allows the two paths to be aligned span by span.

3.6 Arbitration Agent

We align $\mathcal{Y}^E$ and $\mathcal{Y}^S$ by shared span id or by span IoU:

$$\text{IoU}(a, b) = \frac{|[s_a, e_a] \cap [s_b, e_b]|}{|[s_a, e_a] \cup [s_b, e_b]|},$$

using a fixed alignment threshold of 0.5. Arbitration is triggered when one path predicts a mention and the other does not, when aligned mentions assign different entity types, or when overlapping mentions refer to different spans. Low-confidence candidates and singleton predictions without sufficient support are also sent to arbitration. Agreement clusters are accepted directly when their confidence and rationale are sufficient. For triggered cases, the Arbitration Agent selects the supported mention, revises type or span through an existing candidate, or abstains. Singleton candidates are controlled by a separate confidence guard.

$$\text{arb}(k) = \mathbb{K}[d_{\text{exist}} \vee d_{\text{type}} \vee d_{\text{span}} \vee \text{lowconf}].$$

Dataset	# Test	Dataset	# Test
CoNLL2003	3,453	OntoNotes5	8,262
BC5CDR	5,865	Twitter Broad	2,000
NCBI Disease	940	WikiGold	369
MIT Movie Trivia	1,953	WNUT-17	1,287
MIT Restaurant	1,521		

Table 1: Number of test sentences in each evaluation dataset.

Arbitration is local to triggered candidate clusters rather than a full sentence-level re-extraction. Its input contains the aligned candidate predictions, path-specific rationales, confidence scores, and any retrieved or memory support linked to the same span ids. Its output is a single canonical mention decision or abstention for each cluster.

3.7 Deterministic Verification and Memory

Deterministic Verification. Verification is deterministic and uses no LLM. It validates JSON readability, schema membership, offset correctness, and quote alignment, then drops invalid mentions. Here, “verified” means valid under deterministic checks, not gold-correct. A final confidence filter removes low-confidence predictions.

Memory. Memory is initialized per dataset run and updated only after verification, storing reusable entity-level records such as canonical text, normalized key, entity type, dataset/schema, confidence, and seen count (Zhang et al., 2024). During inference, Memory is read before agent prompting and written only after final verification. Records are written, promoted, and retrieved through confidence and frequency thresholds; unpromoted candidate records are not returned for later samples (Zhang et al., 2026). Sample-local identifiers such as seed_0001 are not part of the stored semantic key. By surfacing previously verified entity-level records before Expert retrieval planning, Memory can handle repeated or similar span ambiguities without external lookup. This reduces the number of RAG requests that agents need to issue, which in turn can lower runtime cost and end-to-end latency during inference.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our method on nine widely used NER benchmarks covering general-domain, biomedical, social media, and task-

Method	CoNLL2003	BC5CDR	NCBI Disease	MIT Movie Tri.	MIT Rest.	OntoNotes5	Twitter Broad	WikiGold	WNUT-17	Avg.
<i>DeepSeek-V3.2</i>										
CodeIE	64.87	57.99	49.96	50.86	62.52	58.14	50.07	64.97	42.83	55.80
GPT-NER	62.66	61.26	67.34	55.95	55.24	59.09	55.29	55.31	39.32	56.83
Self-Improving	76.69	70.56	58.17	44.15	52.90	61.54	27.89	74.67	47.07	57.07
Decomposed-QA	70.75	73.83	66.45	51.86	46.42	52.54	53.21	61.13	39.75	57.33
KDR	72.02	71.18	53.14	54.74	60.53	60.51	48.93	74.57	29.81	58.38
ChatIE	67.58	73.31	61.26	55.70	63.50	55.10	40.83	71.09	48.59	59.66
CMAS	77.38	69.68	68.97	54.70	58.33	58.62	57.94	76.84	47.58	63.34
SSDA-NER (ours)	80.65	74.50	69.79	54.79	66.53	62.75	57.95	79.08	53.30	66.59
Expert path	76.26	70.16	58.76	49.19	61.14	55.51	55.10	73.90	48.64	60.96
Structure path	77.17	70.42	57.90	49.55	60.97	55.25	54.80	74.82	47.90	60.98
<i>Qwen3-Max</i>										
Decomposed-QA	68.80	72.53	61.96	54.82	60.13	50.75	54.09	71.21	45.83	60.01
GPT-NER	65.95	73.61	63.06	52.48	56.77	66.14	58.64	64.74	43.10	60.50
Self-Improving	70.95	70.16	64.88	49.73	49.82	64.76	43.63	81.62	53.14	60.97
KDR	76.35	74.33	57.32	57.45	62.44	62.72	52.57	76.82	33.90	61.54
ChatIE	68.41	73.33	68.73	60.03	63.78	53.11	50.70	73.24	50.12	62.38
CodeIE	74.12	70.53	67.34	52.37	68.29	69.11	53.52	66.72	46.17	63.13
CMAS	78.40	74.85	63.79	55.67	59.62	65.83	61.91	77.98	53.59	65.74
SSDA-NER (ours)	80.30	75.11	69.32	60.46	69.41	66.21	60.34	82.24	57.84	69.03
Expert path	76.71	72.76	66.22	53.73	64.82	58.82	54.79	80.89	50.37	64.35
Structure path	75.76	72.18	65.11	57.84	66.44	60.46	56.74	79.15	49.48	64.80
<i>GPT-4o-mini</i>										
GPT-NER	61.82	64.49	50.69	43.55	46.37	54.46	54.03	58.37	31.20	51.66
CodeIE	66.43	63.32	50.10	45.75	56.12	52.66	43.99	68.92	37.86	53.91
Self-Improving	71.95	59.91	52.58	41.18	52.32	54.34	57.37	71.07	39.68	55.60
Decomposed-QA	69.34	67.45	53.03	41.17	48.25	53.33	54.76	66.59	48.55	55.83
KDR	68.90	70.16	50.07	44.77	58.90	56.09	56.79	70.25	32.76	56.52
ChatIE	67.06	64.78	52.61	48.46	55.55	52.28	51.12	67.98	49.32	56.57
CMAS	72.30	65.73	61.74	51.62	57.18	55.72	56.47	74.67	46.72	60.24
SSDA-NER (ours)	74.79	68.83	72.40	54.27	64.74	57.03	58.23	78.78	50.72	64.42
Expert path	68.64	62.44	65.33	46.71	55.20	52.23	55.35	70.33	43.95	57.80
Structure path	70.32	64.75	65.20	47.04	56.35	53.18	55.79	71.10	42.14	58.43

Table 2: F1 comparison across nine datasets and three backbone LLMs. Bold marks the best result for each backbone–dataset setting.

oriented domains: CoNLL2003 (Tjong Kim Sang and De Meulder, 2003), BC5CDR (Li et al., 2016), NCBI Disease (Doğan et al., 2014), MIT Movie Trivia (Ushio and Camacho-Collados, 2021), MIT Restaurant (Ushio and Camacho-Collados, 2021), OntoNotes5 (Weischedel et al., 2013), Twitter Broad (Derczynski et al., 2016), WikiGold (Balasuriya et al., 2009), and WNUT-17 (Derczynski et al., 2017). These datasets cover diverse schemas and phenomena, from standard newswire entities to noisy user text and biomedical-domain mentions.

The number of test sentences for each dataset is summarized in Table 1.

Baselines. We compare with seven non-finetuned LLM-based NER baselines: Self-Improving (Xie et al., 2024), KDR-Agent (Mu et al., 2026), CodeIE (Li et al., 2023), ChatIE (Wei et al., 2024), Decomposed-QA (Xie et al., 2023), GPT-NER (Wang et al., 2025a), and CMAS (Wang et al., 2025b). Except for using the same backbone LLMs as SSDA-NER, baseline configurations follow the corresponding papers and released source code; when multiple official settings are available, we use the best-performing setting for comparison. These methods cover task reformulation, question-answering decomposition, code-style extraction, self-improved demonstrations, knowledge-

enhanced reasoning, and multi-agent collaboration.

Backbone LLMs. To examine whether the conclusions are consistent across different LLM backbones, we conduct experiments with DeepSeek-V3.2 (DeepSeek-AI et al., 2025), Qwen3-Max (Yang et al., 2025), and GPT-4o-mini (OpenAI, 2024). Each baseline and SSDA-NER use the same backbone, dataset split, and evaluation protocol.

Evaluation Metric. We report micro-averaged Precision, Recall, and F1 score. A predicted entity is counted as correct only when both its span and entity type exactly match the gold annotation. Unless otherwise specified, we use micro-F1 as the main metric for comparison.

Implementation Details. The main alignment threshold is IoU 0.5, and the final confidence filter is 0.22. Direct seeds use protected and additional acceptance thresholds of 0.92 and 0.94, respectively. Span augmentation keeps at most 20 added spans, with a minimum confidence of 0.72; Structure-path proposals require 0.90 when transferred as support. RAG uses span-level Wikipedia retrieval with at most three requests per sample, two pages per request, and 420-character summaries. Memory retrieves at most two active records with minimum confidence 0.82 and promotes records

CoNLL-2003				BC5CDR				NCBI Disease			
DeepSeek	E 76	S 166	B 4388	DeepSeek	E 39	S 36	B 7340	DeepSeek	E 23	S 13	B 545
Qwen	E 83	S 153	B 4409	Qwen	E 15	S 151	B 7073	Qwen	E 5	S 11	B 590
GPT-4o	E 93	S 417	B 4134	GPT-4o	E 29	S 106	B 6508	GPT-4o	E 9	S 41	B 627
MIT Movie Trivia				MIT Restaurant				OntoNotes5			
DeepSeek	E 20	S 25	B 3118	DeepSeek	E 35	S 20	B 1782	DeepSeek	E 68	S 87	B 6024
Qwen	E 16	S 537	B 2919	Qwen	E 10	S 116	B 1920	Qwen	E 56	S 118	B 6215
GPT-4o	E 26	S 83	B 2978	GPT-4o	E 30	S 70	B 1588	GPT-4o	E 61	S 95	B 5713
Twitter Broad				WikiGold				WNUT-17			
DeepSeek	E 22	S 17	B 2210	DeepSeek	E 6	S 19	B 590	DeepSeek	E 14	S 15	B 640
Qwen	E 9	S 184	B 2023	Qwen	E 6	S 16	B 614	Qwen	E 16	S 29	B 643
GPT-4o	E 41	S 149	B 2156	GPT-4o	E 8	S 46	B 574	GPT-4o	E 16	S 47	B 663

Figure 3: Exact-match correct spans by path. E, S, and B denote Expert-only, Structure-only, and shared correct spans; correctness requires both offsets and type to match gold.

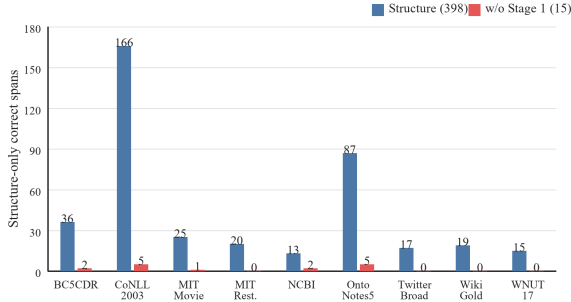


Figure 4: Effect of removing Structure Stage 1 on DeepSeek-V3.2. Bars count Structure-only exact-match correct spans, where both offsets and type match gold.

after three verified mentions within the dataset run.

4.2 Main Results

We evaluate SSDA-NER across domains and backbone LLMs.

Table 2 reports the main results across datasets and backbone LLMs. Against the strongest non-SSDA baseline in each model-dataset setting, SSDA-NER obtains 23 wins, 0 ties, and 4 losses across 27 settings. This supports the claim that SSDA-NER improves in most settings, but not uniformly. The observed losses include DeepSeek-V3.2 on MIT Movie Trivia, Qwen3-Max on OntoNotes5 and Twitter Broad, and GPT-4o-mini on BC5CDR. The gains are especially clear on datasets with noisy expressions or heterogeneous entity schemas, such as WikiGold and WNUT-17. These settings require more than semantic type compatibility: a candidate span must also be judged by how it is positioned and used in the sentence.

The single-path variants further clarify the source of improvement. Both Expert path and Structure path are competitive, but neither dominates the full system. The Expert path tends to

Dataset	Full	w/o Mem.	w/o RAG	w/o Arb.
CoNLL2003	80.65	80.29	78.11	70.45
BC5CDR	74.50	74.42	71.05	65.82
NCBI Disease	69.79	69.42	65.37	56.87
MIT Movie Trivia	54.79	53.38	52.20	47.98
MIT Restaurant	66.53	66.53	64.95	57.74
OntoNotes5	62.75	61.22	58.36	54.56
Twitter Broad	57.95	56.64	55.18	51.06
WikiGold	79.08	78.38	77.78	72.74
WNUT-17	53.30	52.26	49.53	47.26

Table 3: Ablation results of SSDA-NER on nine NER benchmarks with DeepSeek-V3.2.

provide stronger semantic type analysis, while the Structure path supplies complementary local cues about span function. Arbitration combines the two into a more stable final decision process.

4.3 Complementarity Analysis

We then examine whether the Semantic and Structure paths provide complementary predictions.

Figure 3 shows that the two paths are complementary rather than redundant. It counts only exact-match correct spans: both the character offsets and the entity type must match the gold annotation. Across datasets and backbones, most correct spans are shared by both paths, but each path also contributes unique correct spans. The Structure path is not merely a weaker copy of the Expert path: on several datasets, it recovers correct spans that the Expert path misses, especially when the local position and function of a mention provide cues beyond semantic type matching.

This pattern supports the central design of SSDA-NER. If the two paths always made the same mistakes, arbitration would add little value. Instead, the non-overlapping exact-match correct spans indicate that the paths expose different views of the same candidate set. Arbitration can then choose, revise, or reject predictions.



Figure 5: Distribution of conflict types in arbitration.

4.4 Does Structure Guidance Matter?

We further examine whether schema-withheld Structure Stage 1 is necessary for preserving this complementary signal.

Figure 4 isolates the role of Stage 1 in the Structure path. In this ablation, Stage 1 is replaced by the normal direct NER agent because Stage 2 requires selected units as input. With the full Structure path, DeepSeek-V3.2 produces about 398 exact-match correct spans that differ from the Expert path. When Stage 1 is removed, the count drops to about 15. This sharp reduction shows that without the structural analysis introduced by structure guidance, the second path rarely contributes correct spans that are different from the Expert path.

The ablated path also becomes more similar to the model’s pre-extracted seed candidates. In other words, removing Stage 1 leaves little independent signal beyond what the initial extraction already offers. This supports our use of schema-free sentence decomposition before type allocation: Stage 1 makes the Structure path distinct from another semantic pass over seed candidates.

4.5 Ablation Study

We next examine how the main components contribute to the complete framework.

Table 3 evaluates the main components using DeepSeek-V3.2. Removing the Arbitration Agent causes the largest degradation on every dataset, confirming that the main benefit does not come from simply running two paths in parallel. RAG and Memory also help, but their effects are smaller and dataset-dependent, consistent with their roles as auxiliary support and reuse modules. In the w/o Arbitration setting, we directly take the union of the Expert-path and Structure-path outputs. All predictions proposed by either path are retained without resolving existence, type, or span disagreements. The remaining confidence filtering and deterministic verification are unchanged.

Dataset	USD/100	Mem	RAG	LLM	w/	w/o	Save
CoNLL2003	0.310	3076	665	26744	10.71	11.33	0.62
BC5CDR	0.306	4077	5092	49570	18.02	19.72	1.71
NCBI Disease	0.378	833	122	10238	3.66	4.08	0.41
MIT Movie Trivia	0.441	1883	1761	16169	8.75	9.32	0.58
MIT Restaurant	0.341	1373	1542	11702	5.23	5.53	0.30
OntoNotes5	0.294	3304	1672	46840	24.42	25.78	1.35
Twitter Broad	0.289	1766	1983	14986	5.91	6.60	0.70
WikiGold	0.282	218	192	2875	1.09	1.35	0.26
WNUT-17	0.312	1059	1009	9982	4.11	4.54	0.43

Table 4: DeepSeek-V3.2¹ efficiency statistics. USD/100 is cost per 100 sentences; Mem/RAG/LLM are counts; w/ and w/o are full-dataset costs with and without memory.

Method	CoNLL2003	WNUT-17	OntoNotes5
CMAS	67h 58m	42h 54m	158h 12m
SSDA-NER	43h 56m	20h 07m	95h 49m

Table 5: Full-dataset runtime comparison with CMAS.

4.6 Conflict Analysis

Figure 5 summarizes the conflicts passed to arbitration. Existence conflicts dominate, followed by type conflicts and then span conflicts. This pattern shows that the main uncertainty is usually whether a candidate should be accepted as an entity, not whether a span should be mechanically repaired. The Arbitration Agent therefore compares path-specific rationales and confidence scores before making an accept, or reject decision.

4.7 Efficiency Analysis

Finally, we analyze the computational cost and practical efficiency of SSDA-NER.

Table 4 reports the cost and invocation profile of SSDA-NER. The average cost ranges from 0.282 to 0.441 USD per 100 sentences, and the LLM call count makes the multi-agent overhead explicit. Memory reduces total cost on every dataset by reusing verified entity-level records before retrieval planning. Table 5 compares wall-clock runtime with CMAS because both methods decompose NER into multiple LLM-based analysis stages over the complete dataset; SSDA-NER is faster in all three representative datasets in this comparison.

4.8 Case Study

Figure 6 shows how SSDA-NER can recover a span omitted by candidate construction. The Constructor initially provides only *Avis*; Expert and

¹DeepSeek-V3.2 official pricing: Input 2 Yuan / 1M token (0.4 Yuan cache hit); Output 3 Yuan / 1M token. Costs were converted using 1 USD = 6.7387 CNY on 25 August 2026.

Input sentence: Budget said its frequent - flier costs are "substantially below" Avis's level. Gold: Budget ⇒ ORG; Avis ⇒ ORG; Note: no available memory retrieve in this sample	
Constructor: ① "spans": { "span_id": "seed_0001", "start": 69, "end": 73, "text": "Avis" }	Structure agent: Stage 1: "Budget" possibly to be an entity. ② Stage 2: "new_spans": { "start": 0, "end": 6, "text": "Budget", "type_hints": ["ORG"], "confidence": 0.9, "rationale": "Budget is the subject of the verb 'said', indicating it is an entity capable of speech or reporting, which is typical for an organization (ORG). It is being contrasted with 'Avis' in a business context."}]
Expert agent: ② Request retrieve on Avis: Is Avis a PER or ORG? Retriever: Avis Car Rental, LLC is a Global car rental company headquartered in Parsippany, New Jersey, ORG.	NER inference ③ Accept Expert suggestion. Avis, as a car rental company, is appropriately classified as an ORG entity
Arbitration Agent: ④ "span_augmentation": { "enabled": true, "proposals_total": 1, "proposals_used": 1, "added_count": 1, "added_span_ids": ["aug_0001"] }	NER inference ③ Accept Structure suggestion. Avis has been recognised as an ORG entity. Budget has entity structural status; it is recommended that recognize as an ORG entity.

Figure 6: Case study where the Structure path supplements Budget/ORG after incomplete candidate span.

retrieval mainly confirm this existing candidate. In contrast, Structure Stage 1 marks *Budget* as an entity-bearing position because it is the subject of a reporting verb. Stage 2 then assigns ORG with high confidence. The final output recovers both gold ORG mentions. This example illustrates why the Structure path is useful beyond post-processing: it can supply wider candidate before final arbitration; More normalized traces are in Appendix A.2.

5 Conclusion

We presented SSDA-NER, a semantic-structural dual-path framework for LLM-based named entity recognition. It separates semantic type judgment from structural analysis, compares the two path predictions with rationales and confidence scores, and uses arbitration to decide localized disagreements. The results suggest that structural guidance is most useful as a second source of support for arbitration, rather than as a replacement for semantic expertise. Across nine benchmarks and three backbones, SSDA-NER improves in most settings; ablations confirm the value of arbitration in the full system. These findings point to a practical direction for LLM-based NER: keep semantic type judgment, but pair it with independently constrained structural analysis before final arbitration.

Limitations

SSDA-NER relies on multiple calls to LLMs, resulting in higher computational overhead compared to single-prompt approaches. On the evaluated datasets, the measured computational overhead was found to be relatively high, and the number of calls provides additional insight into the sources of this

cost. Although RAG can handle ambiguous entities, its role is limited to retrieving relevant content for candidate snippets, and its effectiveness depends on the ability to retrieve and summarise useful external pages. The memory is updated only through deterministically verified entity-level records; if updated online during test runs, duplicate entities may introduce order sensitivity, which will only stabilise after a certain number of tests. Finally, the structural path provides location and functional analysis guided by LLMs, rather than external symbolic parsing; whilst this mitigates biases associated with single-view extraction, it cannot resolve all ambiguities in fragment selection or annotation.

Acknowledgments

This research work is supported by the National Key Research and Development Program of China under Grant NO. 2024YFF0729003, the National Natural Science Foundation of China under Grant NOs. 62176014, the Fundamental Research Funds for the Central Universities.

References

- Dhananjay Ashok and Zachary C. Lipton. 2023. [Promptner: Prompting for named entity recognition](#). *Preprint*, arXiv:2305.15444.
- Fan Bai, Hamid Hassanzadeh, Ardavan Saeedi, and Mark Dredze. 2025. [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*, pages 28372–28392, Suzhou, China. Association for Computational Linguistics.
- Dominic Balasuriya, Nicky Ringland, Joel Nothman, Tara Murphy, and James R. Curran. 2009. [Named entity recognition in Wikipedia](#). In *Proceedings of the 2009 Workshop on The People’s Web Meets NLP: Collaboratively Constructed Semantic Resources (People’s Web)*, pages 10–18, Suntec, Singapore. Association for Computational Linguistics.
- Siena Biales. 2023. [SAB at SemEval-2023 task 2: Does linguistic information aid in named entity recognition?](#) In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1131–1137, Toronto, Canada. Association for Computational Linguistics.
- DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian

- Yang, and 245 others. 2025. [Deepseek-v3.2: Pushing the frontier of open large language models](#). *Preprint*, arXiv:2512.02556.
- Leon Derczynski, Kalina Bontcheva, and Ian Roberts. 2016. [Broad Twitter corpus: A diverse named entity recognition resource](#). In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1169–1179, Osaka, Japan. The COLING 2016 Organizing Committee.
- Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. [Results of the WNUT2017 shared task on novel and emerging entity recognition](#). In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark. Association for Computational Linguistics.
- Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. [Ncbi disease corpus: A resource for disease name recognition and concept normalization](#). *Journal of Biomedical Informatics*, 47:1–10.
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. [Improving factuality and reasoning in language models through multiagent debate](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 11733–11763. PMLR.
- Besnik Fetahu, Sudipta Kar, Zhiyu Chen, Oleg Rokhlenko, and Shervin Malmasi. 2023. [SemEval-2023 task 2: Fine-grained multilingual named entity recognition \(MultiCoNER 2\)](#). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2247–2265, Toronto, Canada. Association for Computational Linguistics.
- Carlos-Emiliano Gonzalez-Gallardo, Thi Hong Hanh Tran, Nancy Girdhar, Emanuela Boros, Jose G. Moreno, and Antoine Doucet. 2023. [L3I++ at SemEval-2023 task 2: Prompting for multilingual complex named entity recognition](#). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 807–814, Toronto, Canada. Association for Computational Linguistics.
- 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.
- Fei Li, ZhiChao Lin, Meishan Zhang, and Donghong Ji. 2021. [A span-based model for joint overlapped and discontinuous named entity recognition](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers), pages 4814–4828, Online. Association for Computational Linguistics.
- Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wieggers, and Zhiyong Lu. 2016. [Biocreative v cdr task corpus: a resource for chemical disease relation extraction](#). *Database*, 2016:baw068.
- Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2022a. [A survey on deep learning for named entity recognition](#). *IEEE Trans. on Knowl. and Data Eng.*, 34(1):50–70.
- Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2022b. [Unified named entity recognition as word-word relation classification](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10965–10973.
- Peng Li, Tianxiang Sun, Qiong Tang, Hang Yan, Yuanbin Wu, Xuanjing Huang, and Xipeng Qiu. 2023. [CodeIE: Large code generation models are better few-shot information extractors](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15339–15353, Toronto, Canada. Association for Computational Linguistics.
- Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. [Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing](#). *ACM Comput. Surv.*, 55(9).
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. [Self-refine: Iterative refinement with self-feedback](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.
- 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](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(38):32528–32536.
- OpenAI. 2024. [Gpt-4o mini: Advancing cost-efficient intelligence](#). OpenAI Release.
- Yongliang Shen, Xinyin Ma, Zeqi Tan, Shuai Zhang, Wen Wang, and Weiming Lu. 2021. [Locate and label: A two-stage identifier for nested named entity recognition](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2782–2794, Online. Association for Computational Linguistics.

- Zeqi Tan, Shen Huang, Zixia Jia, Jiong Cai, Yinghui Li, Weiming Lu, Yueting Zhuang, Kewei Tu, Pengjun Xie, Fei Huang, and Yong Jiang. 2023. [DAMO-NLP at SemEval-2023 task 2: A unified retrieval-augmented system for multilingual named entity recognition](#). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2014–2028, Toronto, Canada. Association for Computational Linguistics.
- Erik F. Tjong Kim Sang and Fien De Meulder. 2003. [Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition](#). In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.
- Yiqi Tong, Yidong Chen, and Xiaodong Shi. 2021. [A multi-task approach for improving biomedical named entity recognition by incorporating multi-granularity information](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4804–4813.
- Yiqi Tong, Wei Guo, Wei Guo, Fuzhen Zhuang, Deqing Wang, and Huarui Wu. 2026. [Elstar: An elastic long- and short-term multi-interest user modeling network for agricultural news recommendation](#). *Knowledge-Based Systems*, 336:115284.
- Yiqi Tong, Fuzhen Zhuang, Deqing Wang, Haochao Ying, and Binling Wang. 2022. [Improving biomedical named entity recognition with a unified multi-task mrc framework](#). In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8332–8336.
- Asahi Ushio and Jose Camacho-Collados. 2021. [T-NER: An all-round python library for transformer-based named entity recognition](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 53–62, Online. Association for Computational Linguistics.
- Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025a. [GPT-NER: Named entity recognition via large language models](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4257–4275, Albuquerque, New Mexico. Association for Computational Linguistics.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [Self-consistency improves chain of thought reasoning in language models](#). In *The Eleventh International Conference on Learning Representations*.
- Zihan Wang, Ziqi Zhao, Yougang Lyu, Zhumin Chen, Maarten de Rijke, and Zhaochun Ren. 2025b. [A co-operative 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.
- Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nanwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. [Ontonotes release 5.0](#). LDC2013T19.
- Shuhui Wu, Yongliang Shen, Zeqi Tan, and Weiming Lu. 2022. [Propose-and-refine: A two-stage set prediction network for nested named entity recognition](#). In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4418–4424. International Joint Conferences on Artificial Intelligence Organization. Main Track.
- Tingyu Xie, Qi Li, Jian Zhang, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2023. [Empirical study of zero-shot NER with ChatGPT](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7935–7956, Singapore. Association for Computational Linguistics.
- 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.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. [Tree of thoughts: Deliberate problem solving with large language models](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822. Curran Associates, Inc.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023b. [React: Synergizing reasoning and acting in language models](#). In *The Eleventh International Conference on Learning Representations*.
- Fuwei Zhang, Zhao Zhang, Fuzhen Zhuang, Yu Zhao, Deqing Wang, and Hongwei Zheng. 2024. [Temporal knowledge graph reasoning with dynamic memory](#)

enhancement. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):7115–7128.

Fuwei Zhang, Fuzhen Zhuang, Zhao Zhang, and Pengpeng Zhao. 2026. [Temporal knowledge graph reasoning via multi-granularity knowledge refinement](#). *Frontiers of Computer Science*, 20(9):2009374.

A Prompt and Dataset Schemas

A.1 Prompt Schemas

All LLM-facing modules use YAML templates and strict JSON outputs. The excerpts below keep the runtime style while omitting dataset-specific text.

```
prompt_designer:
  system: |
    You design dataset-specific expert
    prompts for schema-driven NER systems.
    Return strict JSON only.
  user: |
    TASK: generate_expert_prompt
    Dataset id: {dataset_id}
    Entity schema: {entity_schema}
    Base expert prompt contracts to
    preserve: {base_expert_prompt}
    Hard requirements:
    - Output design_summary and
    expert_guidance.
    - Generate 8-12 concise bullets about
    mention validity, type semantics,
    ambiguity handling, span-edit caution
    , retrieval triggers, and confidence.
    - Do not include examples, copied
    entity names, fixed keyword lists, or
    dataset-specific lookup rules.
```

```
expert_retrieval_agent:
  system: |
    You are Expert Retrieval Planner.
    Return strict JSON only.
  user: |
    TASK: expert_retrieval_agent
    Text: {text}
    Candidate spans: {candidate_spans}
    Memory top-k: {memory_items}
    Entity schema: {entity_types}
    Hard constraints:
    - Request retrieval only for existing
    span_id values.
    - Request retrieval only for entity
    identity or type disambiguation.
    - Prefer no retrieval when local text
    is sufficient.
    Output JSON:
      retrieval_requests:
        - request_id: str
          span_ids: [str]
          question: str
          priority: high|medium|low
          rationale: str
      global_notes: [str]
      rationale: str
```

```
expert_agent:
  system: |
    You are ExpertAgent.
    Return strict JSON only.
  user: |
    TASK: expert_agent
    Text: {text}
    Candidate spans: {candidate_spans}
    Memory top-k: {memory_items}
    RAG handoff package: {rag_handoff}
```

```
Entity schema: {entity_types}
Span augmentation policy: {
  augmentation_policy}
Hard constraints:
- Use RAG facts only as hints.
- Support quote must be a substring of
  text.
- Unsupported spans receive low
  confidence or are omitted.
Output JSON:
  per_span:
    <span_id>:
      candidate_types: [str]
      excluded_types: [str]
      supporting_quotes: [{quote: str,
        start: int, end: int}]
      confidence: float
      rag_ack: {used_fact_ids: [str], notes
        : str}
      new_spans:
        - {start: int, end: int, text: str,
          type_hints: [str],
            confidence: float,
          supporting_quotes: [object], rationale:
            str}
      rationale: str
```

```
structure_agent:
  system: |
    You are StructureAgent for NER. The
    runtime key is `structure_agent`.
    Return strict JSON only.
  user: |
    TASK:
      structure_agent_as_structure_agent
    Text: {text}
    Candidate spans: {candidate_spans}
    Span augmentation policy: {
      augmentation_policy}
    stage1_input:
      text: {text}
      candidate_spans: {candidate_spans}
      entity_schema: null
      memory_facts: null
      rag_facts: null
      examples: null
    Stage 1 Decompose:
    - Do not use entity type names, label
      descriptions, world knowledge,
      examples, RAG facts, Memory facts, or
      lexical semantic compatibility.
    - Treat content words and named strings
      as opaque units.
    - Use only order, offsets, punctuation,
      apposition, coordination, and
      prepositions as cues for position and
      function.
    stage2_input:
      selected_units: <stage1_decompose.
        selected_units>
      entity_schema: {entity_types}
    Stage 2 Allocate type:
    - Reintroduce the schema only for Stage
      1 selected units.
    Output JSON:
      stage1_decompose:
        schema_used: false
        selected_units:
          - {span_id: str, start: int, end:
```

```

int, text: str,
    structural_role: str,
structural_cues: [str],
    confidence: float, rationale:
str}
stage2_allocate_type:
    schema_used: true
    per_span:
        <span_id>: {candidate_types: [str
], excluded_types: [str],
                confidence: float,
supporting_quotes: [object], rationale:
str}
    new_spans: []
    rationale: str

```

```

ner_expert_agent / ner_structure_agent:
user: |
    TASK: path-level NER inference
    Text: {text}
    Candidate spans: {candidate_spans}
    Expert or Structure guidance: {
constraints}
    Entity schema: {entity_types}
    Output JSON:
        mentions:
            - span_id: str
              ent_type: str
              confidence: float
              supporting_quotes: [{quote: str,
start: int, end: int}]
              rationale: str
              span_ops: []

adjudicator_agent:
user: |
    TASK: adjudicator_agent
    Conflict cluster: {cluster}
    Expert mentions: {expert_mentions}
    Structure mentions: {context_mentions}
    Output JSON:
        winner_source: expert|structure|none
        winner_span_id: str
        winner_ent_type: str
        confidence: float
        rationale: str

verifier:
    deterministic_checks:
        - JSON readability
        - schema membership
        - exact offset alignment
        - quote alignment
    memory_record:
        canonical_text: str
        ent_type: str
        confidence: float
        seen_count: int
        provenance: object

```

A.2 Case Studies of Path Disagreement

We screened 15 path-disagreement cases and selected three that best illustrate type, existence, and task-oriented structure disagreements. The traces below are condensed and format-normalized to

match the agent schemas in this paper; Stage 1 fields are reconstructed from the Structure path protocol and observed path outputs.

```

[
    "conll2003/test-00200", "conll2003/test-00214",
    "conll2003/test-00318", "conll2003/test-00319",
    "conll2003/test-00699", "conll2003/test-00944",
    "conll2003/test-01599", "ontonotes5_en/test-01682",
    "ontonotes5_en/test-00548", "wikigold/696",
    "mit_restaurant/test-00996", "mit_restaurant/test-00735",
    "mit_restaurant/test-01276", "mit_restaurant/test-00388",
    "twitter_broad/test-01179"
]

```

Type Disagreement: CoNLL2003 test-00200.

```

{
    "input": {
        "dataset": "CoNLL2003",
        "sample_id": "test-00200",
        "text": "League duties restricted the Barbarians' selectorial options but they still boast 13 internationals including England full-back Tim Stimpson and recalled wing Tony Underwood, plus All Black forwards Ian Jones and Norm Hewitt."
    },
    "gold_mentions": [
        {"text": "Barbarians", "start": 29, "end": 39, "ent_type": "ORG"},
        {"text": "England", "start": 110, "end": 117, "ent_type": "LOC"},
        {"text": "Tim Stimpson", "start": 128, "end": 140, "ent_type": "PER"},
        {"text": "Tony Underwood", "start": 159, "end": 173, "ent_type": "PER"},
        {"text": "All Black", "start": 180, "end": 189, "ent_type": "ORG"},
        {"text": "Ian Jones", "start": 199, "end": 208, "ent_type": "PER"},
        {"text": "Norm Hewitt", "start": 213, "end": 224, "ent_type": "PER"}
    ],
    "candidate_pre_extraction": {
        "source": "candidate_agent + direct_seed",
        "fallback_added": false,
        "spans": [
            {"span_id": "seed_0001", "text": "Barbarians", "start": 29, "end": 39},
            {"span_id": "seed_0006", "text": "England", "start": 110, "end": 117},
            {"span_id": "seed_0007", "text": "All Black", "start": 180, "end": 189}
        ],
        "note": "Direct-seed candidates exist, so no generic low-confidence fallback spans are added."
    },
}

```

```

"expert_agent": {
  "per_span": {
    "seed_0006": {
      "candidate_types": ["MISC"],
      "excluded_types": ["PER", "ORG"],
      "confidence": 0.85,
      "supporting_quotes": [{"quote": "
England", "start": 110, "end": 117}],

      "rationale": "Semantic reading
treats England as a national
modifier in a sports phrase."
    },
    "seed_0007": {
      "candidate_types": ["MISC"],
      "excluded_types": ["PER", "LOC"],
      "confidence": 0.90,
      "supporting_quotes": [{"quote": "
All Black", "start": 180, "end":
189}],
      "rationale": "Semantic reading
treats All Black as a team-related
descriptor."
    }
  },
  "structure_agent": {
    "stage1_decompose": {
      "schema_used": false,
      "selected_units": [
        {"span_id": "seed_0006", "role":
"pre-head modifier of player role",
"function": "affiliation/location
cue"},
        {"span_id": "seed_0007", "role":
"pre-head modifier of plural player
role", "function": "team/group cue"}
      ],
      "rationale": "Both units occur
before sport-role heads and license
entity-bearing positions before type
names are considered."
    },
    "stage2_allocate_type": {
      "per_span": {
        "seed_0006": {
          "candidate_types": ["LOC"],
          "confidence": 0.85,
          "rationale": "Position before
full-back marks geographic
affiliation for the person mention."
        },
        "seed_0007": {
          "candidate_types": ["ORG"],
          "confidence": 0.80,
          "rationale": "Position before
forwards marks an organized team or
squad name."
        }
      }
    }
  },
  "path_predictions": {
    "expert": [
      {"span_id": "seed_0006", "ent_type":
"MISC"},
      {"span_id": "seed_0007", "ent_type":
"MISC"}
    ],
    "structure": [

```

```

      {"span_id": "seed_0006", "ent_type":
"LOC"},
      {"span_id": "seed_0007", "ent_type":
"ORG"}
    ]
  },
  "arbitration_agent": {
    "triggered_by": ["type"],
    "decision": [
      {"span_id": "seed_0006", "
winner_source": "structure", "
winner_ent_type": "LOC"},
      {"span_id": "seed_0007", "
winner_source": "structure", "
winner_ent_type": "ORG"}
    ],
    "rationale": "The structural roles
explain why the same strings
function as typed mentions rather
than generic descriptors."
  },
  "final_mentions": [
    {"text": "England", "start": 110, "
end": 117, "ent_type": "LOC"},
    {"text": "All Black", "start": 180, "
end": 189, "ent_type": "ORG"}
  ]
}

```

Existence Disagreement: OntoNotes5 test-01682.

```

{
  "input": {
    "dataset": "OntoNotes5",
    "sample_id": "test-01682",
    "text": "According to a news report
on February 14th by Xinhua News
Agency from Beijing , when the multi-
national forces carried out an air
attack on Baghdad on the 13th ,
hundreds of civilians were killed in
the bombing ."
  },
  "gold_mentions": [
    {"text": "February 14th", "start":
30, "end": 43, "ent_type": "DATE"},
    {"text": "Xinhua News Agency", "start":
47, "end": 65, "ent_type": "ORG"},

    {"text": "Beijing", "start": 71, "end":
78, "ent_type": "GPE"},
    {"text": "Baghdad", "start": 141, "
end": 148, "ent_type": "GPE"},
    {"text": "the 13th", "start": 152, "
end": 160, "ent_type": "DATE"},
    {"text": "hundreds", "start": 163, "
end": 171, "ent_type": "CARDINAL"}
  ],
  "candidate_pre_extraction": {
    "source": "candidate_agent +
direct_seed",
    "fallback_added": false,
    "spans": [
      {"span_id": "seed_0001", "text": "
February 14th", "start": 30, "end":
43},
      {"span_id": "seed_0002", "text": "
Xinhua News Agency", "start": 47, "
end": 65},

```

```

    {"span_id": "seed_0003", "text": "
    Beijing", "start": 71, "end": 78},
    {"span_id": "seed_0005", "text": "
    Baghdad", "start": 141, "end": 148},
    {"span_id": "seed_0004", "text": "
    the 13th", "start": 152, "end": 160},

    {"span_id": "seed_0006", "text": "
    hundreds", "start": 163, "end": 171}
  ],
},
"expert_agent": {
  "path_output": [],
  "rationale": "The semantic path does
  not emit final mentions for this
  sample after candidate-level routing
  ."
},
"structure_agent": {
  "stage1_decompose": {
    "schema_used": false,
    "selected_units": [
      {"span_id": "seed_0001", "role":
      "object of preposition on in report
      frame"},
      {"span_id": "seed_0002", "role":
      "agent of by in source frame"},
      {"span_id": "seed_0003", "role":
      "complement of from in source-
      location frame"},
      {"span_id": "seed_0005", "role":
      "complement of on in attack frame"},
      {"span_id": "seed_0004", "role":
      "temporal complement of on after
      attack"},
      {"span_id": "seed_0006", "role":
      "quantity head before plural
      civilians"}
    ],
    "rationale": "The sentence contains
    a report-source frame followed by
    an attack-event frame; selected
    units occupy canonical date, source,
    location, and quantity positions."
  },
  "stage2_allocate_type": {
    "per_span": {
      "seed_0001": {"candidate_types":
      ["DATE"], "confidence": 0.95},
      "seed_0002": {"candidate_types":
      ["ORG"], "confidence": 0.95},
      "seed_0003": {"candidate_types":
      ["GPE"], "confidence": 0.90},
      "seed_0005": {"candidate_types":
      ["GPE"], "confidence": 0.90},
      "seed_0004": {"candidate_types":
      ["DATE"], "confidence": 0.90},
      "seed_0006": {"candidate_types":
      ["CARDINAL"], "confidence": 0.90}
    }
  }
},
"path_predictions": {
  "expert": [],
  "structure": [
    {"span_id": "seed_0001", "ent_type":
    "DATE"},
    {"span_id": "seed_0002", "ent_type":
    "ORG"},
    {"span_id": "seed_0003", "ent_type":

```

```

    "GPE"},
    {"span_id": "seed_0005", "ent_type":
    "GPE"},
    {"span_id": "seed_0004", "ent_type":
    "DATE"},
    {"span_id": "seed_0006", "ent_type":
    "CARDINAL"}
  ]
},
"arbitration_agent": {
  "triggered_by": ["existence"],
  "winner_source": "structure",
  "rationale": "The structure path
  provides aligned position and
  function rationales for all six seed
  spans, while the expert path
  abstains."
},
"final_mentions": [
  {"text": "February 14th", "start":
  30, "end": 43, "ent_type": "DATE"},
  {"text": "Xinhua News Agency", "start":
  47, "end": 65, "ent_type": "ORG"},

  {"text": "Beijing", "start": 71, "end":
  78, "ent_type": "GPE"},
  {"text": "Baghdad", "start": 141, "end":
  148, "ent_type": "GPE"},
  {"text": "the 13th", "start": 152, "end":
  160, "ent_type": "DATE"},
  {"text": "hundreds", "start": 163, "end":
  171, "ent_type": "CARDINAL"}
]
}

```

Task-Oriented Structure: MIT Restaurant test-00996.

```

{
  "input": {
    "dataset": "MIT Restaurant",
    "sample_id": "test-00996",
    "text": "make me a reservation for a
    restaurant with a nonsmoking patio
    and that isnt fancy for 30 minutes
    from now"
  },
  "gold_mentions": [
    {"text": "reservation", "start": 10,
    "end": 21, "ent_type": "Amenity"},
    {"text": "nonsmoking patio", "start":
    46, "end": 62, "ent_type": "Amenity"},
    {"text": "isnt fancy", "start": 72, "end":
    82, "ent_type": "Amenity"},
    {"text": "30 minutes from now", "start":
    87, "end": 106, "ent_type": "Hours"}
  ],
  "candidate_pre_extraction": {
    "source": "candidate_agent +
    direct_seed",
    "fallback_added": false,
    "spans": [
      {"span_id": "seed_0001", "text": "
      nonsmoking patio", "start": 46, "end":
      62},
      {"span_id": "cand_0001", "text": "
      restaurant", "start": 32, "end": 43}
    ],

```

```

    "note": "The existing seed prevents
generic fallback expansion."
},
"expert_agent": {
  "per_span": {
    "seed_0001": {
      "candidate_types": ["Amenity"],
      "confidence": 0.90,
      "supporting_quotes": [{"quote": "
nonsmoking patio", "start": 46, "end
": 62}]
    },
    "cand_0001": {
      "candidate_types": ["
Restaurant_Name"],
      "confidence": 0.70,
      "rationale": "restaurant is
semantically related to the task but
is generic."
    }
  },
  "path_output": []
},
"structure_agent": {
  "stage1_decompose": {
    "schema_used": false,
    "selected_units": [
      {"span_id": "seed_0001", "role":
"complement of with", "function": "
facility or seating feature"},
      {"span_id": "cand_0001", "role":
"generic search head", "function": "
not an entity-bearing name"}
    ],
    "rationale": "The unit after with
modifies the requested restaurant as
an available facility; the generic
head restaurant is not selected as a
named entity."
  },
  "stage2_allocate_type": {
    "per_span": {
      "seed_0001": {
        "candidate_types": ["Amenity"],
        "excluded_types": ["Rating", "
Location", "Restaurant_Name", "Price
", "Hours", "Dish", "Cuisine"],
        "confidence": 0.95,
        "rationale": "The span names a
seating restriction/facility offered
by the restaurant."
      }
    }
  },
  "path_predictions": {
    "expert": [],
    "structure": [
      {"span_id": "seed_0001", "ent_type
": "Amenity", "confidence": 0.95}
    ]
  },
  "arbitration_agent": {
    "triggered_by": ["existence"],
    "winner_source": "structure",
    "winner_span_id": "seed_0001",
    "winner_ent_type": "Amenity",
    "rationale": "The structure path
supplies a local function rationale
for the with-prepositional phrase."
  }
}

```

```

    },
    "final_mentions": [
      {"text": "nonsmoking patio", "start":
46, "end": 62, "ent_type": "Amenity"}
    ]
  }
}

```

A.3 Dataset Entity Type Schemas

CoNLL2003. PER: A named individual person, including full names and personal name mentions.

ORG: A named organization, institution, company, team, government body, or other formal collective entity.

LOC: A named location such as a country, city, region, or other geographic place.

MISC: A named entity that does not fit PER, ORG, or LOC, such as nationalities, events, works, products, or other proper named categories.

BC5CDR. Chemical: A chemical, drug, compound, or other named chemical substance mentioned in biomedical text. Disease: A disease, disorder, syndrome, adverse condition, or other pathological condition mentioned in biomedical text.

NCBI Disease. Disease: A disease, disorder, syndrome, abnormal condition, or disease-class mention annotated as a biomedical disease entity.

MIT Movie Trivia. Actor: A named actor or performer associated with a movie.

Plot: A plot, storyline, topic, or narrative description used to identify or describe a movie.

Opinion: An evaluative or subjective phrase expressing opinion about a movie.

Award: A named award, prize, honor, or award category associated with movies.

Year: A year or release-time expression associated with a movie.

Genre: A movie genre or style category.

Origin: A country, region, language, or cultural origin associated with a movie.

Director: A named director associated with a movie.

Soundtrack: A soundtrack, song, music cue, or named musical item associated with a movie.

Character_Name: A named fictional character in a movie.

Quote: A quoted line, dialogue fragment, or memorable phrase associated with a movie.

MIT Restaurant. Rating: A restaurant quality descriptor such as best, popular, top-rated, or similar rating-oriented language.

Amenity: A restaurant amenity, service, seating feature, reservation option, delivery option, or other offered facility.

Location: A place reference for restaurant search, including nearby, downtown, city areas, neighborhoods, or directional location phrases.

Restaurant_Name: The proper name of a restaurant, cafe, chain, or dining establishment.

Price: A price descriptor such as cheap, inexpensive, expensive, or other cost-related expression.

Hours: A time or opening-hours expression describing when a restaurant is open or available.

Dish: A dish, meal item, menu item, or food item requested or described by the user.

Cuisine: A cuisine style or food category such as mexican, italian, sushi, vegan, or similar dining type.

OntoNotes5. CARDINAL: A numeral or quantity that does not fall into another specialized numeric category.
 DATE: An explicit date expression such as a day, month, year, or calendar date span.
 PERSON: A named person or individual human.
 NORP: A nationality, religious group, or political group.
 GPE: A geopolitical entity such as a country, city, state, or government-associated place.
 LAW: A named law, legal case, treaty, regulation, or legislative act.
 PERCENT: A percentage expression.
 ORDINAL: An ordinal number or ranked position.
 MONEY: A monetary amount including currency expressions.
 WORK_OF_ART: A titled work of art such as a book, song, film, painting, or other creative work.
 FAC: A facility, building, airport, highway, bridge, or other constructed physical infrastructure.
 TIME: A time expression smaller than a full date, such as a clock time or part of day.
 QUANTITY: A measured amount such as weight, distance, volume, or similar quantity.
 PRODUCT: A named product, object, vehicle, device, or commercial item.
 LANGUAGE: A named human language.
 ORG: An organization such as a company, institution, agency, or named group.
 LOC: A non-geopolitical location such as a mountain range, body of water, or other place name.
 EVENT: A named event such as a war, sports event, festival, or other occurrence.

Twitter Broad. PER: Mentions of individual people, persons, or named human identities in tweets.
 ORG: Mentions of organizations, institutions, companies, teams, agencies, or other named collective entities in tweets.
 LOC: Mentions of locations, places, geographic regions, facilities, or other named locative entities in tweets.

WikiGold. LOC: A geographic or geopolitical location such as a city, country, region, landmark, or other place name.
 MISC: A miscellaneous named entity that does not fit person, organization, or location, such as events, nationalities, works, or other proper-name categories.
 ORG: An organization such as a company, institution, government body, sports team, or other formally named group.
 PER: A person name, including full names, surnames, given names, or named individuals.

WNUT-17. corporation: A named company, business, brand owner, or other corporate organization.
 creative-work: A named creative work such as a song, movie, book, show, game, or other titled artistic or media work.
 group: A named group, team, band, organization-like collective, or community that is not best described as a corporation.
 location: A named location such as a city, country, region, venue, or other geographic place.
 person: A named individual person, including full names, handles, or user references that identify a person.
 product: A named product, device, service offering, application, vehicle model, or other commercial or technical product name.

B Additional Analysis

B.1 Threshold Selection and Deterministic Fusion

Most thresholds control routing, candidate propagation, and LLM-call cost rather than acting as

performance-optimized test-set decision boundaries. They were validated with all three backbones on CoNLL2003 during system development. For span augmentation, we first tested 0.5, 0.6, 0.7, and 0.8, then refined the range with 0.72, 0.74, and 0.76. Precision and Recall remained comparable in the 0.7–0.8 range and on CoNLL2003 and NCBI Disease; we therefore selected 0.72. The high Structure-path transfer threshold prioritizes reliable supplementation and reduces false positives.

For deterministic confidence-based fusion, identical predictions are accepted directly, conflicting predictions are resolved by selecting the higher-confidence prediction, and singleton predictions are retained only when confidence is at least 0.82. On CoNLL2003 with DeepSeek-V3.2 this baseline obtains 76.30 F1, with 4,527 true positives, 1,692 false positives, and 1,121 false negatives.