DR-DEE: Demonstration Retrieval, Schema Pruning, and Structured Decoding for Chinese Document-Level Event Extraction

Anonymous ACL submission

Abstract

Chinese document-level event extraction (DEE) is dominated by supervised graph and prompt-based models trained from scratch on each benchmark. Frozen large language models (LLMs) are an attractive alternative, but published augmentations are typically validated on a single benchmark and a bespoke pipeline, obscuring what actually matters. We present DR-DEE, an inference pipeline that wraps three deterministic steps around one frozen-LLM call: demonstration retrieval, event-type retrieval for schemas with more than thirty types, and structured decoding against a typed schema. Without task-specific training, DR-DEE reaches 90.0 / 73.8 / 43.2 / 43.4 Arg-C F_1 on CHFIN, DuEE-Fin, CMNEE, and DEIE: it surpasses every reported supervised and LLM-augmentation baseline on CHFIN, is within 0.4 F_1 of the closest supervised graph model on DuEE-Fin, and trails the supervised ceiling by 26 and 9 F_1 on CMNEE and DEIE. Within-pipeline ablations attribute the entire $+27.7$ F_1 gain over zero-shot to demonstration retrieval; the other two components mainly stabilize the longest schema. Replacing single DR-DEE hooks with five published LLM augmentations on Qwen3-14B yields above-noise gains on at most one dataset. Error analysis traces the residual on CMNEE and DEIE to span-boundary drift rather than hallucination, identifying span-level repair as the next direction.

1 Introduction

Document-level event extraction (DEE) maps a document and a typed schema to a list of typed events, each with role-filled arguments (Yang et al., 2018; Zheng et al., 2019; Zhu et al., 2024; Ren et al., 2024). Compared to sentence-level extraction, DEE requires reasoning over arguments scattered across paragraphs and frequently shared across multiple events. Chinese DEE in particular is dominated by supervised systems—graph neural networks, parallel-record decoders, and schema-aware prompts trained from scratch on each benchmark (Xu et al., 2021; Yang et al., 2021; Liang et al., 2022; Wang et al., 2023a; Pan et al., 2024; Xu et al., 2024; Ren et al., 2025; Liu et al., 2025)—which incur a high per-benchmark training cost and have slowed in progress on the harder trigger-annotated schemas (CMNEE, DEIE).

Frozen large language models (LLMs) handle long contexts and declarative schemas at inference time, motivating prompting as a training-free alternative. A growing literature stacks augmentations on top of frozen LLMs—retrieval-augmented generation, in-context learning, self-consistency, role-knowledge prompts, contrastive demonstrations, two-step decomposition, and verifier-based decoding, among others (Wang et al., 2023b; Hu et al., 2023; Mo et al., 2024; Liu and Luo, 2024; Manakul et al., 2023; Liang et al., 2025; Zhang et al., 2026)—but each is typically validated on a single benchmark and a bespoke implementation, leaving open how a minimal frozen-LLM pipeline for Chinese DEE should be assembled and which augmentations actually carry its gain.

We design DR-DEE, a single-call inference pipeline for Chinese DEE that wraps three deterministic steps around one frozen-LLM call: dense demonstration retrieval, dense event-type retrieval for long ontologies, and structured decoding with typed-schema validation. Each step targets a distinct zero-shot failure mode—missing annotation conventions, prompt overload on large schemas, and malformed JSON output—and can be ablated in place. We evaluate DR-DEE at full scale on four Chinese DEE benchmarks (CHFIN, DuEE-Fin, CMNEE, DEIE), spanning 5–64 event types and both no-trigger and explicit-trigger conventions, and re-implement five representative augmentations as drop-in replacements of single DR-DEE hooks on Qwen3-14B for matched comparison.

Our main contributions are:

A single-call pipeline for Chinese DEE. DR-DEE wraps three deterministic steps around one frozen-LLM call per document: dense demonstration retrieval over labeled training documents, dense event-type retrieval that prunes long ontologies, and structured decoding with typed-schema validation. Each component targets a distinct zero-shot failure mode and can be ablated independently.

Strong performance without task-specific training. On four Chinese DEE benchmarks at full scale, DR-DEE surpasses every reported supervised and LLM-augmentation baseline on CHFIN, is on par with the closest recent supervised graph model on DuEE-Fin, and reduces but does not close the supervised gap on CMNEE and DEIE (Table 1).

Demonstration retrieval drives the gain over zero-shot. A within-pipeline ablation attributes the entire average gain over zero-shot to demonstration retrieval; event-type retrieval and structured decoding are near-neutral on average and pay off only on the longest schema, where they reduce prompt-length-induced format drift (Table 2).

Published prompt augmentations seldom help on Qwen3-14B. Of five representative augmentations re-implemented as drop-in replacements of single DR-DEE hooks, none yields an above-noise gain on more than one of the four datasets. The LLM-as-Judge verifier even degrades F_1 on three of four datasets, because the dominant residual error is span-boundary drift rather than hallucination—a failure mode verifier prompts cannot directly target.

The remaining gap is at the span level. An argument-level error analysis attributes the residual on CMNEE and DEIE to span-boundary drift (correct role, wrong span boundaries) rather than type or hallucinated-argument errors, identifying span-level repair as the most promising next direction without re-introducing per-dataset training.

We will release the DR-DEE pipeline and the four-dataset full-scale evaluation harness to support reproducible study of frozen-LLM Chinese DEE.

2 Related Work

Supervised document-level event extraction. Document-level event extraction was formulated by Yang et al. (2018) and Zheng et al. (2019) on Chinese financial announcements, and has since developed along two complementary axes. *Decoding* relaxes the fixed generation order of sequential decoders via parallel-record models (Yang et al., 2021; Huang et al., 2023). *Structural priors* encode sentence–mention or token–token graphs, relation-augmented attention, proxy nodes, and schema- or type-aware decoding (Xu et al., 2021; Liang et al., 2022; Wang et al., 2023a; Wan et al., 2023; Pan et al., 2024; Xu et al., 2024; Zhao et al., 2024), with recent variants based on dynamic hypergraphs, dynamic proxies, BiLSTM-enhanced GNNs, and pattern–instance graphs (Ren et al., 2025; Zhao et al., 2026; Liu et al., 2025; Wan et al., 2025; Liao et al., 2026). All are trained from scratch on each benchmark; we use them as the supervised reference and ask how close a frozen-LLM pipeline can get without task-specific training.

Argument-only and event-aware extraction. A parallel line of work assumes triggers or event types are given and studies argument filling alone, including conditional generation (Lu et al., 2021; Hsu et al., 2022), two-step head detection (Zhang et al., 2020), AMR-aware encoders (Xu et al., 2022), optimal transport (Pouran Ben Veyseh et al., 2022), event-aware context modeling (Zeng et al., 2022), chain-style reasoning (Liu et al., 2023), semantic prompt extractors (Zhou et al., 2025), dual-paradigm fusion (Yu et al., 2025), and role pattern–instance graphs (Wan et al., 2025). This literature consistently identifies span-boundary drift and role confusion as the dominant failure modes, an observation we corroborate within DR-DEE (§7).

LLMs and prompting for event extraction. Frozen-LLM augmentations for event extraction span self-consistency (Wang et al., 2023b), role-knowledge prompts (Hu et al., 2023), contrastive in-context learning (Mo et al., 2024), definition-driven and two-step extraction (Zhang et al., 2020; Liu and Luo, 2024; Liang et al., 2025), and LLM-as-Judge verification (Manakul et al., 2023). For Chinese DEE specifically, schema-aware adaptive RAG (Liang et al., 2025) and active structure learning over GLM (Zhang et al., 2026) report strong results on individual benchmarks. Each augmentation is typically validated on a single benchmark and a bespoke implementation, and is rarely shipped as a drop-in component on top of a shared pipeline. We therefore re-implement five representative augmentations as drop-in replacements of single DR-DEE hooks (§3.6).

Position of this work. The closest prior pipelines retrieve schema or document context for a frozen

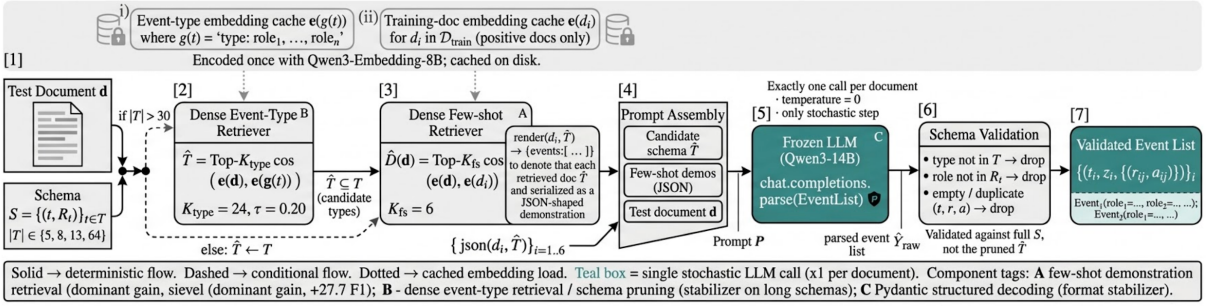

Figure 1: The DR-DEE pipeline. Steps (i)–(iii) and (v) are deterministic; only step (iv) calls the LLM. e is the dense encoder; $g(t)$ is the textual signature of type t; embeddings are cached. Schema pruning activates only when $|\mathcal{T}|>30$.

LLM (Liang et al., 2025). DR-DEE differs in three ways: (i) the retrieval unit is a labeled training document rendered as a JSON-shaped demonstration rather than schema descriptions or unlabeled text; (ii) the pipeline issues one LLM call per document and is otherwise deterministic, which makes component-level ablations tractable; and (iii) we evaluate at full scale on four Chinese benchmarks spanning no-trigger and explicit-trigger conventions, financial and non-financial domains, and 5–64 event types.

3 Method

3.1 Task Formulation

We treat document-level event extraction as a structured prediction task with no task-specific training. A document d is a sequence of sentences and a schema $\mathcal{S} = \{(t, \mathcal{R}_t)\}_{t\in\mathcal{T}}$ assigns to each event type $t \in \mathcal{T}$ a set of admissible roles $\mathcal{R}_t$. An extractor produces a list of events $\{(t_i, z_i, \{(r_{ij}, a_{ij})\}_j)\}_i$, where z_i is an optional trigger and each argument pairs a role $r_{ij} \in \mathcal{R}_{t_i}$ with a span a_{ij} in d. We report event-type detection F_1 (Event), argument identification F_1 (Arg-I; span only), and the strict argument classification metric F_1 (Arg-C; both span and role).

3.2 Pipeline Overview

DR-DEE is a single-call retrieval-augmented inference pipeline (Figure 1). For a test document d, the pipeline executes five steps; only step (4) calls the LLM, and the remaining four are deterministic:

1. **Schema pruning**: when $|\mathcal{T}|$ exceeds a threshold, retrieve a top-K candidate set $\hat{\mathcal{T}} \subseteq \mathcal{T}$ (§3.4).
2. **Demonstration retrieval**: retrieve a small set of training documents similar to d and render each as a JSON-shaped demonstration filtered to $\hat{\mathcal{T}}$ (§3.3).
3. **Prompt assembly**: build a single prompt containing the candidate schema, the retrieved demonstrations, and d.
4. **LLM call**: invoke the frozen LLM with structured decoding against a typed event-list schema (§3.5).
5. **Schema validation**: re-validate the parsed output against the original $\mathcal{S}$, dropping types and roles outside the whitelist as well as empty or duplicate arguments.

The three components target three distinct zero-shot failure modes: demonstration retrieval transmits annotation conventions, schema pruning curbs format drift on long ontologies, and structured decoding eliminates malformed JSON. The pipeline order is fixed so that each component can be ablated independently (§6). All retrieval cost is amortized into one-time, cached embedding computation; per-document inference involves exactly one LLM call regardless of schema size. We use Qwen3-14B (Yang et al., 2025) as the LLM and Qwen3-Embedding-8B as the dense encoder e.

3.3 Dense Demonstration Retrieval

A central observation in our error analysis is that zero-shot LLMs make systematic mistakes about the *annotation conventions* of each benchmark, not about the underlying event semantics: which mentions count as arguments, whether triggers are head words or full verb phrases, and how multi-event documents are partitioned. Retrieved demonstrations directly transmit these conventions from labeled training data to the model.

We encode every positive training document d_i once with the dense encoder e, and at inference

time retrieve the K_{fs} most similar training documents to d by cosine similarity:

$$\hat{\mathcal{D}}(d) = \underset{d_i \in \mathcal{D}_{\text{train}}}{\text{Top-}K_{fs}} \cos\big(e(d), e(d_i)\big). \quad (1)$$

Each retrieved d_i is rendered as a JSON-shaped demonstration `{events: [. . . ]}` after filtering its annotated events by $\hat{\mathcal{T}}$, so demonstrations cannot reintroduce types already pruned by the schema retriever. We restrict the retrieval pool to documents with at least one annotated event, and use $K_{fs} = 6$ (chosen on the CHFIN validation split). Embeddings are cached, so demonstration retrieval adds no LLM cost at inference time.

3.4 Dense Event-Type Retrieval

Long ontologies inflate prompt length, which causes format drift in long-context LLMs, and expose the model to many distractor types that share roles with the correct one. We address both with a contextual schema pruner. For each $t \in \mathcal{T}$, we encode a textual signature

$$g(t) = \texttt{type:}\ \text{role}_1, \ldots, \text{role}_n, \quad (2)$$

once with e and rank types by $\cos(e(d), e(g(t)))$ at inference time. The prompt retains the top K_{type} types whose similarity exceeds τ, with a hard floor of five for short documents. The retriever activates only when $|\mathcal{T}| > 30$; for the smaller schemas in our cohort it is a no-op. We use $K_{type} = 24$ and $\tau = 0.20$. Validation in step (5) uses the original $\mathcal{S}$, so a correct type recovered by the LLM is never discarded by the retriever.

3.5 Structured Decoding and Validation

We declare a typed schema with fields `events`, `event_type`, `trigger`, `arguments`, `role`, and `argument`, and decode through a constrained-decoding interface that rejects outputs which violate the schema. The parsed output is then revalidated against the original $\mathcal{S}$: events whose type lies outside $\mathcal{T}$, arguments whose role lies outside $\mathcal{R}_t$ for the predicted type, empty argument strings, and exact duplicate (t, r, a) triples are removed. Validating against $\mathcal{S}$ rather than $\hat{\mathcal{T}}$ ensures that a correct type recovered by the LLM is never discarded by an upstream retrieval decision, and separates the failure modes of the type retriever and the LLM for ablation.

3.6 Drop-in Augmentations from Prior Work

To position DR-DEE against prior LLM-prompting work under matched conditions—same backbone, retrieval pool, and decoder—we re-implement five representative augmentations as drop-in replacements of either the LLM call (step 4) or post-processing (steps 5–6). Each is evaluated as a separate row in Table 1; we never stack augmentations.

- **Self-Consistency** (Wang et al., 2023b): sample $n = 3$ outputs at temperature 0.4 and keep (*event_type*, *role*, *argument*) tuples receiving at least two votes.
- **Role hints** (Hu et al., 2023): derive role-knowledge prompts from training co-occurrence statistics and prepend them to the schema block.
- **Negative demonstrations** (Mo et al., 2024): inject up to one empty-event training document into the demonstration block.
- **Coarse-to-fine** (Zhang et al., 2020; Liu and Luo, 2024): ask the LLM for high-level event families derived from the prefix of each type label, and supplement $\hat{\mathcal{T}}$ with up to 12 additional types from the selected families before the main extraction call.
- **Verifier** (Manakul et al., 2023): issue a second LLM call to judge each predicted argument triple, and drop triples whose probability falls below 0.5.

4 Experimental Setup

4.1 Datasets

We evaluate at full scale on four Chinese DEE benchmarks: CHFIN (Zheng et al., 2019), DuEE-Fin under the local-split protocol (Li et al., 2020; Han et al., 2022), CMNEE (Zhu et al., 2024), and DEIE (Ren et al., 2024). They span 5–64 event types, no-trigger and explicit-trigger conventions, and four domains. Following Shaar et al. (2025), who show that trigger conventions are not directly comparable across benchmarks, we report each dataset's canonical argument metric: overall record-level F_1 on CHFIN and DuEE-Fin, argument F_1 on CMNEE standard test, and Arg-C on DEIE. Detailed statistics, splits, and metric definitions are in Appendix A.

4.2 Backbone and Hyperparameters

All experiments use Qwen3-14B (Yang et al., 2025) through an OpenAI-compatible endpoint

System	Mode	CHFIN			DuEE-Fin			CMNEE			DEIE		
		P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Supervised baselines (graph and prompt-based)													
DCFEE-O (Yang et al., 2018)	sup.	68.0	63.3	65.6	59.8	55.5	57.6	30.3	22.3	25.7	–	–	–
Doc2EDAG (Zheng et al., 2019)	sup.	82.7	75.2	78.8	67.1	60.1	63.4	54.3	23.9	33.2	–	–	–
DE-PPN (Yang et al., 2021)	sup.	83.7	76.4	79.9	69.0	33.5	45.1	38.2	35.0	36.5	–	–	–
GIT (Xu et al., 2021)	sup.	83.6	76.9	80.1	69.8	65.9	67.8	–	–	35.8	–	–	–
RAAT (Liang et al., 2022)	sup.	82.9	79.3	81.1	69.2	57.4	62.8	–	–	–	–	–	–
PTPCG (Wan et al., 2023)	sup.	83.2	74.9	78.8	71.0	61.7	66.0	–	–	–	–	–	–
ProCNet (Wang et al., 2023a)	sup.	84.4	80.9	82.7	78.6	72.6	75.5	–	–	–	–	–	–
SEA (Zhao et al., 2024)	sup.	–	–	80.2	–	–	–	–	–	–	–	–	–
CAINet (Pan et al., 2024)	sup.	84.3	82.9	83.6	79.1	72.9	75.8	–	–	–	–	–	–
SEELE (Xu et al., 2024)	sup.	–	–	85.1	–	–	80.8	–	–	–	–	–	–
EEQA (Zhu et al., 2024; Ren et al., 2024)	sup.	–	–	–	–	–	–	39.0	39.1	39.0	53.5	30.2	38.6
Text2Event (Zhu et al., 2024)	sup.	–	–	–	–	–	–	31.3	41.3	35.5	–	–	–
BERT+CRF (Zhu et al., 2024; Ren et al., 2024)	sup.	–	–	–	–	–	–	63.1	52.3	57.2	46.2	30.0	36.4
GATED (Cui et al., 2024)	sup.	–	–	–	–	–	–	59.9	56.5	58.2	–	–	–
BART-Gen (Ren et al., 2024)	sup.	–	–	–	–	–	–	–	–	–	54.3	43.5	48.3
PAIE (Ren et al., 2024; Zhu et al., 2024)	sup.	–	–	–	–	–	–	72.0	67.0	**69.4**	58.8	47.2	**52.4**
Recent supervised baselines (graph variants)													
DSH (Ren et al., 2025)	sup.	84.8	79.9	82.3	78.2	70.6	74.2	–	–	–	–	–	–
DP-SAAM (Zhao et al., 2026)	sup.	–	–	83.4	–	–	76.3	–	–	–	–	–	–
Multi-BiGNN (Liu et al., 2025)	sup.	86.1	82.6	**84.3**	80.1	73.5	76.0	–	–	–	–	–	–
LLM-augmentation baselines (literature, copied verbatim)													
DESIA-CHAT (ChatGPT) (Yu and Cao, 2024)	ICL/RAG	–	–	80.5	–	–	–	–	–	–	–	–	–
GLM-ASLS (Zhang et al., 2026)	ASLS-SFT	86.5	91.4	88.9	82.7	80.9	**81.8**	–	–	–	–	–	–
Frozen Qwen3-14B baseline													
ZS (no retrieval, ours)	zero-shot	59.9	56.0	57.9	45.7	45.5	45.6	17.7	27.8	21.6	20.5	20.5	20.5
DR-DEE (ours, full)	retrieval	88.5	91.6	**90.0**	75.3	72.5	**73.8**	36.5	53.0	**43.2**	42.3	44.7	**43.4**
Published LLM augmentations re-implemented on the DR-DEE backbone (100-doc P1 subsample)													
DR-DEE (P1, no augment, ours)	retrieval	90.7	91.3	**90.9**	69.4	70.9	70.1	43.5	49.7	**46.4**	47.9	50.2	49.0
DR-DEE + Self-Cons. (Wang et al., 2023b)	retrieval+SC	89.5	91.5	90.5	70.6	71.5	**71.0**	44.9	50.5	**47.5**	48.0	50.4	49.2
DR-DEE + RoleHints (Hu et al., 2023)	retrieval+prompt	90.6	90.7	90.7	68.0	69.7	68.9	42.7	49.7	45.9	47.5	49.0	48.2
DR-DEE + NegDemos (Mo et al., 2024)	retrieval+CICL	90.6	91.0	90.8	70.3	71.6	**71.0**	42.1	45.9	43.9	48.0	50.2	49.1
DR-DEE + 2-step (Liu and Luo, 2024)	retrieval+2-step	90.6	91.0	90.8	70.9	67.4	69.1	44.3	41.3	42.8	48.5	52.5	**50.4**
DR-DEE + Verifier (Manakul et al., 2023)	retrieval+judge	91.6	85.2	88.3	75.2	66.4	70.5	48.5	35.1	40.8	51.5	44.7	47.9

Table 1: Main results: F_1 on each dataset's canonical argument metric. Middle block: full-scale ZS and DR-DEE. Bottom block: drop-in augmentations on the DR-DEE backbone, evaluated on a 100-document P1 subsample with DR-DEE (no augment) as reference. Bold marks the best F_1 within each cohort.

and Qwen3-Embedding-8B for retrieval. Decoding temperature is 0 (or 0.4 inside Self-Consistency). Retrieval hyperparameters: $K_{fs} = 6$, $K_{type} = 24$, type-pruning similarity threshold 0.20. Sample-level worker concurrency is 16.

4.3 Comparison Cohorts

We compare DR-DEE against (a) supervised baselines trained on each dataset's official split (DCFEE through PAIE; recent variants DSH, DP-SAAM, Multi-BiGNN), and (b) LLM-augmentation baselines, both copied from the literature (DESIA-CHAT, GLM-ASLS) and re-implemented in our pipeline (§3.6). All literature numbers are copied verbatim from the cited papers.

5 Main Results

Table 1 reports F_1 on each dataset's canonical argument metric. Supervised numbers and the two literature LLM-augmentation rows (DESIA-CHAT, GLM-ASLS) are copied verbatim from the cited papers; the bottom block re-implements five drop-in augmentations on our Qwen3-14B backbone.

Strong performance without task-specific training. DR-DEE reaches 90.0 F_1 on CHFIN, exceeding every supervised baseline (Multi-BiGNN, 84.3) and the strongest LLM-augmentation result on this dataset (GLM-ASLS, 88.9). On DuEE-Fin, it reaches 73.8 F_1, -0.4 F_1 short of DSH (74.2) and 1.7–7.0 F_1 below the rest of the recent supervised graph cluster (CAINet 75.8, ProCNet 75.5, Multi-BiGNN 76.0, DP-SAAM 76.3, SEELE 80.8) and 8.0 F_1 below GLM-ASLS (81.8). On CMNEE and DEIE, the supervised state of the art is set by PAIE (69.4, 52.4); DR-DEE trails by 26.2 and 9.0 F_1 but still surpasses every pre-PAIE supervised baseline except BERT+CRF and GATED on CMNEE and BART-Gen on DEIE. The retrieval-free baseline (ZS) sits 22–32 F_1 below DR-DEE across the four datasets.

Drop-in augmentations seldom help on Qwen3-14B. The bottom block of Table 1 is the most apples-to-apples comparison: every row uses the

Configuration	CHFIN	DuEE	CMNEE	DEIE	Avg. Δ
DR-DEE (full)	90.95	70.13	46.38	49.03	—
− demonstration retrieval (= ZS)	57.89 (−33.06)	45.61 (−24.52)	21.64 (−24.74)	20.47 (−28.56)	−27.72
− event-type retrieval	90.80 (−0.15)	71.23 (+1.10)	45.55 (−0.83)	48.63 (−0.40)	−0.07
− structured decoding	89.82 (−1.13)	71.21 (+1.08)	46.35 (−0.03)	47.52 (−1.51)	−0.40

Table 2: Component ablation on Arg-C F_1 (P1 100-doc subsample); parentheses give deltas vs. the full pipeline.

same Qwen3-14B backbone and modifies a single DR-DEE hook. No augmentation produces a gain larger than 1.5 F_1 on more than one dataset. The two clear wins are Self-Consistency on CMNEE (+1.1) and two-step decoding on DEIE (+1.4). RoleHints is non-positive on every dataset; NegDemos posts small positives on DuEE-Fin and DEIE but loses −2.5 F_1 on CMNEE; the Verifier degrades F_1 on three of four datasets, most severely on CMNEE, where recall drops from 49.7 to 35.1 as the judge deletes correct arguments alongside hallucinated ones. We do not extrapolate this ranking to other backbones (§Limitations).

6 Ablation Study

We disable each DR-DEE component in turn on a stratified 100-document subsample (**P1**) drawn uniformly across event types per dataset, with all other settings held fixed. "− demonstration retrieval" degrades the pipeline to ZS; "− event-type retrieval" passes the full $\mathcal{S}$ to the LLM; "− structured decoding" replaces typed-schema parsing with free-form JSON generation.

Demonstration retrieval drives the gain over zero-shot. Removing it costs 27.7 F_1 on average and is the only intervention that hurts every dataset, with drops of −24.5 to −33.1 F_1. The pipeline collapses to ZS on every split, so the entire +27.7 F_1 gain is attributable to demonstration retrieval. The pattern is invariant to schema size, trigger convention, and domain, suggesting that on Qwen3-14B the bottleneck for zero-shot Chinese DEE lies in the implicit annotation conventions transmitted by labeled training documents rather than the underlying event semantics.

Schema pruning and structured decoding stabilize long ontologies. Their average deltas are −0.07 and −0.40 F_1; on smaller schemas they are even slightly positive, since the type retriever is a no-op and the JSON constraint is unnecessary. On DEIE ($|\mathcal{T}|$=64), removing structured decoding costs 1.5 F_1 due to occasional malformed JSON. A K_{type} sweep on DEIE shows a +6.5 F_1 jump from K_{type}=12 to 24 with diminishing returns thereafter; setting K_{type}=$|\mathcal{T}|$ costs an additional 0.4 F_1, indicating that an unbounded prompt is worse than a bounded but generous one. We therefore keep both components on by default.

Dataset	TP	FP	FN	Mismatch
CHFIN	21,581	2,102	1,240	705
DuEE-Fin	5,098	860	1,081	815
CMNEE	9,098	10,288	3,300	5,548
DEIE	10,471	8,244	7,336	6,064

Table 3: Argument-level error states from DR-DEE. *Mismatch* = correct role with wrong span boundary.

7 Error Analysis

We assign each predicted argument to one of four states: *TP* (correct type, role, span), *FP* (no matching gold), *FN* (missed gold), and *Mismatch* (correct role, wrong span boundary). Mismatches are not captured by hallucination (FP) or omission (FN) and require span-level rather than prompt-side correction.

Span-boundary drift dominates on trigger-annotated schemas. On the financial datasets, Mismatch is small (3.3% and 16.0% of TP) and the residual is dominated by FP and FN. On CMNEE and DEIE, Mismatch reaches 61% and 58% of TP, so span-boundary drift is the dominant residual error. This explains why semantic prompt augmentations help only at the margin: role hints and the Verifier act on the prompt layer while the residual lives at the span layer; the Verifier's recall collapse on CMNEE (49.7→35.1, Table 1) follows directly, since it cannot separate boundary mismatches from hallucinations and drops both.

Errors concentrate on long documents and role-rich types. Beyond the aggregate split in Table 3, two within-dataset axes carry signal (Figure 2). *Document length* (panel a): on the financial cluster the error rate roughly doubles between short and long documents—CHFIN goes from 5.1% at

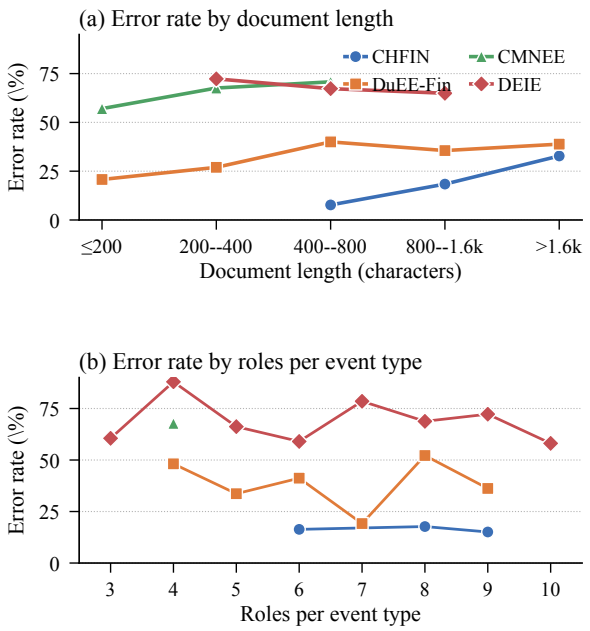

Figure 2: Argument error rate ($1 - \text{Arg-C F}_1$) stratified by document length (a) and by the number of admissible roles in the predicted event type (b). Bins with fewer than 50 arguments are dropped.

400–800 chars to 21.9% at 1.6k+, DuEE-Fin from 19.7% to 29.9% across the same shift—while CMNEE and DEIE sit on a much higher plateau (55–60%) that is largely independent of length, since their documents cluster within a single bin. *Schema width* (panel b): within DEIE, types with 7–9 admissible roles are 5–15 points harder than the easier 5–6-role buckets in the same schema, even though all DEIE types share the same length budget; the financial datasets show no within-dataset slope on this axis. Together these two views identify long inputs and role-rich event types as the regions where the next round of pipeline work is most likely to yield gain—and both axes are orthogonal to the prompt-side augmentations we tested.

Augmentations help only on the dataset whose error mode they target. Self-Consistency exposes alternative entity–event attachments on CMNEE's dense military events (+1.1); two-step decoding mitigates fine-grained subtype confusion within DEIE's 64-type schema (+1.4); on CHFIN and DuEE-Fin, where retrieval already absorbs annotation conventions and the residual is gold-annotation-driven boundary truncation, no augmentation produces an above-noise positive cell. Closing the CMNEE/DEIE gap therefore requires span-level intervention rather than further prompt engineering.

8 Discussion

Three observations follow from §6–7 and inform when this kind of pipeline is worth choosing. First, the gain over zero-shot is concentrated in a single step: demonstration retrieval transmits dataset-specific annotation conventions that the LLM cannot infer from a schema alone. Second, the supervised gap on CHFIN and DuEE-Fin closes without training, but on CMNEE and DEIE it persists because the residual error budget shifts from semantic to span-level—a layer that prompt augmentations cannot reach. Third, the ranking of drop-in augmentations holds only for Qwen3-14B; transfer across LLMs is not in scope.

A practical implication is that boundary-aware repair, e.g. a span aligner trained on a small per-dataset annotation budget, is a lower-cost path than further prompt engineering for closing the CMNEE/DEIE gap. We leave this and cross-LLM validation to future work.

9 Conclusion

DR-DEE is a single-call frozen-LLM pipeline for Chinese DEE that combines demonstration retrieval, event-type retrieval, and structured decoding. Without fine-tuning, it surpasses every reported supervised and LLM-augmentation baseline on CHFIN, comes within 0.4 F_1 of the closest supervised graph model on DuEE-Fin, and trails the supervised ceiling by 9 and 26 F_1 on DEIE and CMNEE. Within-pipeline ablations attribute the entire gain over zero-shot to demonstration retrieval; five drop-in augmentations on Qwen3-14B yield above-noise gains on at most one dataset. Span-boundary drift dominates the residual on CMNEE and DEIE, identifying span-level repair as the next direction.

Limitations

DR-DEE has four practical limitations. (i) All numbers use a single backbone (Qwen3-14B / Qwen3-Embedding-8B); the augmentation ranking in Table 1 should not be extrapolated to other LLMs, and cross-backbone validation is left for future work. (ii) Retrieval replaces task-specific training, so quality degrades when the labeled pool is small, noisy, or out-of-distribution—a regime we do not stress-test. (iii) The dominant residual on CMNEE and DEIE is span-boundary mismatch (Table 3), which none of DR-DEE's components targets. (iv) The pipeline issues one LLM call per document, so

full-scale runs take several wall-clock hours per dataset.

Ethics Statement

We use only publicly available event extraction datasets, accessed under their stated licenses or research-use terms. The pipeline is intended for research on structured information extraction; it is not appropriate for high-stakes downstream decisions without human review. Several of our datasets touch sensitive domains (finance, military news, legal/emergency reporting); deployments should add provenance, dataset-specific validation, and human oversight before use.

References

Yachao Cui, Hongli Yu, Han Cao, Kaiguang Wang, and Yangchun Wang. 2024. A document-level event extraction method based on heterogeneous graphs and gated fusion. In *Proceedings of the IEEE International Conference on Data Mining and Big Data*. IEEE.

Cuiyun Han, Jinchuan Zhang, Xinyu Li, Guojin Xu, Weihua Peng, and Zengfeng Zeng. 2022. DuEE-fin: A large-scale dataset for document-level event extraction. In *Natural Language Processing and Chinese Computing*, pages 172–183, Cham. Springer International Publishing.

I-Hung Hsu, Kuan-Hao Huang, Elizabeth Boschee, Scott Miller, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2022. DEGREE: A data-efficient generation-based event extraction model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1890–1908, Seattle, United States. Association for Computational Linguistics.

Ruijuan Hu, Haiyan Liu, and Huijuan Zhou. 2023. Role knowledge prompting for document-level event argument extraction. *Applied Sciences*, 13(5):3041.

Guanhua Huang, Runxin Xu, Ying Zeng, Jiaze Chen, Zhouwang Yang, and Weinan E. 2023. An iteratively parallel generation method with the pre-filling strategy for document-level event extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10834–10852, Singapore. Association for Computational Linguistics.

Xiaoya Li, Fan Li, Lei Pan, Yuxin Chen, Weihua Peng, Qian Wang, Yajuan Lyu, and Yong Zhu. 2020. DuEE: A large-scale dataset for Chinese event extraction in real-world scenarios. In *Natural Language Processing and Chinese Computing*, pages 534–545. Springer.

Sheng Liang, Hang Lv, Zhihao Wen, Yaxiong Wu, Yongyue Zhang, Hao Wang, and Yong Liu. 2025. Adaptive schema-aware event extraction with retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 7927–7946, Suzhou, China. Association for Computational Linguistics.

Yuan Liang, Zhuoxuan Jiang, Di Yin, and Bo Ren. 2022. RAAT: Relation-augmented attention transformer for relation modeling in document-level event extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4985–4997, Seattle, United States. Association for Computational Linguistics.

Junru Liao, Haoran Wang, Weimin Cheng, and Yatian Shen. 2026. CL4DEE: A continual learning framework for document-level event extraction with mixture-of-experts. *Expert Systems with Applications*, 314:131723.

Jian Liu, Chen Liang, Jinan Xu, Haoyan Liu, and Zhe Zhao. 2023. Document-level event argument extraction with a chain reasoning paradigm. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9570–9583, Toronto, Canada. Association for Computational Linguistics.

Xi Liu, Jun Long, WenTi Huang, Xi Yu, and Yang Yang. 2025. BiLSTM-enhanced GNN for document-level event extraction with multi-strategy optimization. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Zhuoyuan Liu and Yilin Luo. 2024. Document-level event extraction with definition-driven ICL. *Preprint*, arXiv:2408.05566.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction. 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 2795–2806, Online. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

Ying Mo, Jian Yang, Jiahao Liu, Shun Zhang, Jingang Wang, and Zhoujun Li. 2024. C-ICL: Contrastive in-context learning for information extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics.

Bangze Pan, Yang Li, Suge Wang, Xiaoli Li, Deyu Li, Jian Liao, and Jianxing Zheng. 2024. Document-level event extraction via information interaction based on event relation and argument correlation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5156–5166, Torino, Italia. ELRA and ICCL.

Amir Pouran Ben Veyseh, Minh Van Nguyen, Franck Dernoncourt, Bonan Min, and Thien Nguyen. 2022. Document-level event argument extraction via optimal transport. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1648–1658, Dublin, Ireland. Association for Computational Linguistics.

Qi Ren, Weihua Wang, Jie Yu, and Guanglai Gao. 2025. Dynamic structure hypergraph for document-level event extraction. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5. IEEE.

Yubing Ren, Yanan Cao, Hao Li, Yingjie Li, Zixuan ZM Ma, Fang Fang, Ping Guo, and Wei Ma. 2024. DEIE: Benchmarking document-level event information extraction with a large-scale Chinese news dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4592–4604, Torino, Italia. ELRA and ICCL.

Shaden Shaar, Wayne Chen, Maitreyi Chatterjee, Barry Wang, Wenting Zhao, and Claire Cardie. 2025. Are triggers needed for document-level event extraction? *Transactions of the Association for Computational Linguistics*, 13:1560–1577.

Qizhi Wan, Tao Liu, Changxuan Wan, Rong Hu, Keli Xiao, and Yuxin Shuai. 2025. Event pattern-instance graph: A multi-round role representation learning strategy for document-level event argument extraction. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1865–1877, Vienna, Austria. Association for Computational Linguistics.

Qizhi Wan, Changxuan Wan, Keli Xiao, Dexi Liu, Chenliang Li, Bolong Zheng, Xiping Liu, and Rong Hu. 2023. Joint document-level event extraction via token-token bidirectional event completed graph. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10481–10492, Toronto, Canada. Association for Computational Linguistics.

Xinyu Wang, Lin Gui, and Yulan He. 2023a. Document-level multi-event extraction with event proxy nodes and hausdorff distance minimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10118–10133, Toronto, Canada. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*.

Runxin Xu, Tianyu Liu, Lei Li, and Baobao Chang. 2021. Document-level event extraction via heterogeneous graph-based interaction model with a tracker. 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 3533–3546, Online. Association for Computational Linguistics.

Runxin Xu, Peiyi Wang, Tianyu Liu, Shuang Zeng, Baobao Chang, and Zhifang Sui. 2022. A two-stream AMR-enhanced model for document-level event argument extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5025–5036, Seattle, United States. Association for Computational Linguistics.

Zijie Xu, Peng Wang, Wenjun Ke, Guozheng Li, Jiajun Liu, Ke Ji, Xiye Chen, and Chenxiao Wu. 2024. Incorporating schema-aware description into document-level event extraction. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 6597–6605.

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. *arXiv preprint arXiv:2505.09388*.

Hang Yang, Yubo Chen, Kang Liu, Yang Xiao, and Jun Zhao. 2018. DCFEE: A document-level Chinese financial event extraction system based on automatically labeled training data. In *Proceedings of ACL 2018, System Demonstrations*, pages 50–55, Melbourne, Australia. Association for Computational Linguistics.

Hang Yang, Dianbo Sui, Yubo Chen, Kang Liu, Jun Zhao, and Taifeng Wang. 2021. Document-level event extraction via parallel prediction networks. 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 6298–6308, Online. Association for Computational Linguistics.

Dong Yu and JiaWen Cao. 2024. Document-level overlapping event extraction via ChatGPT leading decision and cross-sentence entity representation enhancements. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Hongli Yu, Kaiguang Wang, Han Cao, and Yachao Cui. 2025. Document-level event argument extraction

with dual-paradigm fusion strategy. *Journal of King Saud University Computer and Information Sciences*, 37(10):322.

Qi Zeng, Qiusi Zhan, and Heng Ji. 2022. EA^{2}E: Improving consistency with event awareness for document-level argument extraction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2649–2655, Seattle, United States. Association for Computational Linguistics.

Jinming Zhang, Yanping Chen, Anqi Zou, Ruizhang Huang, and Yongbin Qin. 2026. An active structure-learning strategy for LLMs-based document-level event extraction. In *PRICAI 2025: Trends in Artificial Intelligence*, Lecture Notes in Computer Science, pages 656–663, Singapore. Springer Nature Singapore.

Zhisong Zhang, Xiang Kong, Zhengzhong Liu, Xuezhe Ma, and Eduard Hovy. 2020. A two-step approach for implicit event argument detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7479–7485, Online. Association for Computational Linguistics.

Gang Zhao, Yidong Shi, Shudong Lu, Xinjie Yang, Guanting Dong, Jian Xu, Xiaocheng Gong, and Si Li. 2024. Type-aware decoding via explicitly aggregating event information for document-level event extraction. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024*, pages 11061–11065, Seoul, Republic of Korea. IEEE.

Guoliang Zhao, Shi Li, and Meng Xu. 2026. DP-SAAM: Dynamic proxy nodes and structure-aware adaptive matching for document-level event extraction. In *Pattern Recognition and Computer Vision (PRCV 2025)*, Lecture Notes in Computer Science, pages 435–449, Singapore. Springer Nature Singapore.

Shun Zheng, Wei Cao, Wei Xu, and Jiang Bian. 2019. Doc2EDAG: An end-to-end document-level framework for Chinese financial event extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 337–346, Hong Kong, China. Association for Computational Linguistics.

Yipeng Zhou, Jiaxin Fan, Qingchuan Zhang, Lin Zhu, and Xingchen Sun. 2025. Modeling semantic-aware prompt-based argument extractor in documents. *Applied Sciences*, 15(10):5279.

Mengna Zhu, Zijie Xu, Kaisheng Zeng, Kaiming Xiao, Mao Wang, Wenjun Ke, and Hongbin Huang. 2024. CMNEE: A large-scale document-level event extraction dataset based on open-source Chinese military news. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

A Dataset Details

Table 4 summarizes split sizes and per-dataset evaluation conventions for the four Chinese document-level event extraction benchmarks used in this paper.

Dataset	Domain / language	$\lvert\mathcal{T}\rvert$	Train	Eval
CHFIN	Chinese finance, no-trigger	5	25,632	3,204
DuEE-Fin	Chinese open-domain news	13	6,139	1,171
CMNEE	Chinese military news	8	12,000	2,727
DEIE	Chinese news (legal/emergency)	64	15,987	2,017

Table 4: Datasets used in this paper. $\lvert\mathcal{T}\rvert$: number of event types; "Eval" columns are split sizes.

CHFIN (Zheng et al., 2019). A 5-type, no-trigger Chinese financial event extraction schema collected from the China Financial Annual Announcements corpus. Each record has an event type and a fixed set of role-filled arguments, but no annotated trigger word. We evaluate on the official test split of 3,204 documents using the canonical overall record-level F_1 (Doc2EDAG protocol).

DuEE-Fin (Li et al., 2020; Han et al., 2022). A 13-type Chinese financial benchmark with explicit trigger annotations, drawn from real-world financial news. The public test labels are not released; we therefore follow the literature convention of evaluating on the validation split (1,171 documents) under the author-local-split protocol used by ProCNet, CAINet, DSH, DP-SAAM, Multi-BiGNN, and GLM-ASLS in our comparison cohort. The reported metric is overall record-level F_1.

CMNEE (Zhu et al., 2024). An 8-type Chinese military news schema with 11 argument role types and both event-detection and argument annotations. We use the standard test split (2,727 documents) and report event-detection F_1 and argument F_1. The argument metric is the standard-test-set protocol; we explicitly do not include the Golden-Trigger or Golden-Span variants from the literature, because they condition on extra annotated evidence and are not directly comparable to our end-to-end pipeline.

DEIE (Ren et al., 2024). A 64-type fine-grained Chinese news schema covering legal, emergency, and policy events. Each event has an explicit trigger and several typed argument roles. We use the test split (2,017 documents) and report Arg-I (span match only) and Arg-C (both span and role) following the original benchmark definition. The Head-C metric defined in the original benchmark requires head-word matching; we do not run a head matcher because our pipeline outputs full spans, so Head-C cells are not reported.

Schema sizes and pruning. The pipeline's event-type retriever activates only when $\lvert\mathcal{T}\rvert > 30$. For CHFIN ($\lvert\mathcal{T}\rvert$=5), DuEE-Fin ($\lvert\mathcal{T}\rvert$=13), and CMNEE ($\lvert\mathcal{T}\rvert$=8) it is a no-op and the full schema is always passed to the LLM. Only DEIE ($\lvert\mathcal{T}\rvert$=64) triggers the retriever, with default top-K_{type}=24.

Trigger conventions across datasets. The four datasets differ on triggers: CHFIN has no triggers (events are records); DuEE-Fin, CMNEE, and DEIE all have explicit triggers in the gold annotation. Our pipeline always asks the model for a trigger field; for CHFIN we ignore the trigger at evaluation time. The role of triggers in document-level extraction is itself dataset-dependent (Shaar et al., 2025), which is one reason we report dataset-canonical argument metrics rather than a single uniform metric across all four.