SMADE-IE: Sparse Multi-Agent Framework with Evidence-Driven Debate for Zero-Shot Information Extraction

Anonymous ACL submission

Abstract

Zero-shot information extraction (IE) with large language models (LLMs) has attracted increasing attention due to its flexibility in adapting to new schemas and domains without task-specific training. Existing approaches mainly rely on monolithic prompting, each-type prompting, or multi-agent debate. However, monolithic prompting often suffers from boundary and type errors, while each-type prompting and multi-agent debate introduce cross-type conflicts, redundant agent interactions, and substantial token overhead. To address these challenges, we propose SMADE-IE, a sparse and evidence-driven multi-agent framework for zero-shot IE. SMADE-IE first employs an Adaptive Mode Selector to dynamically route inputs into either a lightweight Global Extraction Mode or a Type-Centric Extraction Mode, reducing unnecessary type selection and reasoning noise. For conflicting predictions, we further introduce an Evidence-Driven Debate mechanism that structures arguments into Toulmin-style components and performs confidence aggregation through external evidence scoring and Bayesian updates. Experimental results on 9 benchmark datasets across NER, RE, and JERE tasks show that SMADE-IE consistently outperforms existing zero-shot IE baselines while also improving token efficiency through sparse agent selection and early-stopping debate.

1 Introduction

Information Extraction (IE) aims to extract structured facts from unstructured text (Li et al., 2023b), supporting applications such as knowledge graph construction and question answering (Liu et al., 2020; Jain, 2020). Representative tasks include named entity recognition (NER), relation extraction (RE), and joint entity and relation extraction (JERE) (Zeng et al., 2014; Zheng et al., 2017). Although supervised and instruction-tuned IE methods achieve strong performance (Lu et al., 2022; Lou et al., 2023), they often rely on extensive annotations, task-specific tuning, or fixed schemas, limiting adaptability to evolving entity types, relation patterns, and domains (Sainz et al., 2024; Fuente et al., 2025).

Some studies leverage the extensive background knowledge of LLMs for zero-shot IE, enabling flexible adaptation to new schemas and domains (Xu et al., 2024). Early LLM-based zero-shot IE methods typically use a *monolithic-prompt* strategy, extracting all elements in a single stage (Xie et al., 2023; Li et al., 2024). Despite its efficiency, this strategy requires jointly predicting entity spans and types, which often leads to missed mentions, boundary errors, and type confusion. As shown in Figure 1(a), the model misses "John Smith", expands "California" to "in California", and misclassifies "Apple" as PRODUCT. To alleviate these issues, recent work has introduced *each-type prompting* (Li et al., 2023a, 2024), which extracts each entity or relation type independently using separate prompts. However, independently generated type-specific predictions can overlap or conflict with one another. For example, in Figure 1(b), "Apple" is simultaneously assigned the types ORGANIZATION and PRODUCT.

To mitigate such conflicts, recent *multi-agent debate* methods employ specialized agents for different entity or relation types and use inter-agent discussions to resolve conflicting predictions (Guan et al., 2025; Lu et al., 2025). However, exhaustive agent interactions still limit their efficiency and reliability. First, since each sample typically contains only a small subset of all entity or relation types, as shown in Table 1, selecting all types for every sample introduces unnecessary token overhead and computational cost. Second, incorrect candidate types inject irrelevant context into the debate, distracting agents from meaningful evidence and leading to unreliable decisions (Fan et al., 2026; Shi et al., 2023; Liu et al., 2024; Yang et al., 2025;

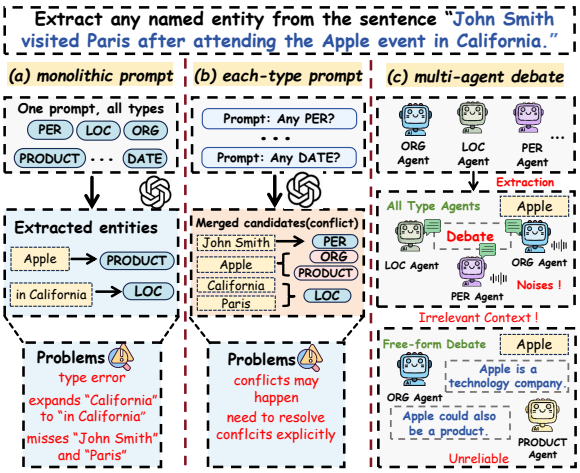

Figure 1: Existing LLM-based zero-shot IE methods.

Task	Dataset	Ontology Size	Avg. Types / Sample
NER	OntoNotes5	18	1.31
	CrossRE	38	4.72
RE	DocRED	32	2.46
	REDFM	32	2.77
	SciERC	8	1.88

Table 1: Statistics of some zero-shot IE benchmarks.

Tian et al., 2026). For instance, as shown in Figure 1(c), three agents (*PER*, *ORG*, and *LOC*) label the same entity with different types, causing redundant discussions that obscure useful evidence and increase the risk of incorrect adjudication. Finally, free-form debates lack explicit structures for modeling claims, points of contention, and supporting evidence, making it difficult to systematically aggregate confidence scores and reach stable decisions (Choi et al., 2026). As illustrated in Figure 1 (c), the PRODUCT agent argues that "Apple could also be a product," while the ORGANIZATION agent claims that "Apple is a technology company." Without a structured mechanism to distinguish reliable evidence from noisy or irrelevant arguments, such debates may fail to converge and incur substantial token overhead (Choi et al., 2026).

To address these challenges, we propose SMADE-IE, a sparse and evidence-driven multi-agent framework for zero-shot IE. As shown in Figure 2, SMADE-IE balances extraction reliability and computational cost through two components. First, the Adaptive Mode Selector estimates sample complexity and selects only relevant types: simple inputs follow the lightweight Global Extraction Mode, while complex inputs enter the Type-Centric Extraction Mode, reducing unnecessary token overhead and irrelevant reasoning noise. Second, for type conflicts in the Type-Centric Extraction Mode, SMADE-IE introduces a structured debate mechanism. Unlike free-form debate, it decomposes candidate predictions into Toulmin-style argumentative components (Gupta et al., 2024), allowing agents to attack the evidence and reasoning behind each claim. An external evidence scorer and Bayesian Beta updates then aggregate support and refutation signals into evidence-centered confidence estimates for final adjudication. The main contributions are as follows:

- We present SMADE-IE, a sparse multi-agent framework for zero-shot information extraction. With the Adaptive Mode Selector, it automatically dispatches samples to either the Global Extraction Mode or the Type-Centric Extraction Mode based on their complexity, balancing efficiency and performance.
- We introduce the Evidence-Driven Debate mechanism that structures each conflict into Toulmin-style components, and combines external evidence scoring with Bayesian Beta updates for more reliable multi-agent conflict adjudication.
- We evaluate SMADE-IE on 9 zero-shot IE datasets and show its advantages in both extraction performance and token efficiency.

2 Methodology

In this section, we first introduce the definition of zero-shot IE tasks and then describe SMADE-IE in detail. As shown in Figure 2, the Adaptive Mode Selector dispatches each input to either the lightweight Global Extraction Mode or the finer-grained Type-Centric Extraction Mode. For cross-type conflicts in the Type-Centric Extraction Mode, we introduce the Evidence-Driven Debate module for structured conflict adjudication. For JERE, an Iterative Entity–Relation Alignment step further enforces ontology consistency between entity and relation predictions.

2.1 Problem Formulation

Given a sentence s and predefined sets of entity types $\mathcal{T}_e$ and relation types $\mathcal{T}_r$, NER aims to identify all entity mentions $\{e_1, \ldots, e_n\}$ in s and assign a type $t_e^i \in \mathcal{T}_e$ to each, producing a set of typed entities $E = \{(e_i, t_e^i)\}$. RE requires first identifying the entity mentions $\{e_1, \ldots, e_n\}$ and then predicting, for each ordered pair of entities (e_i, e_j), the

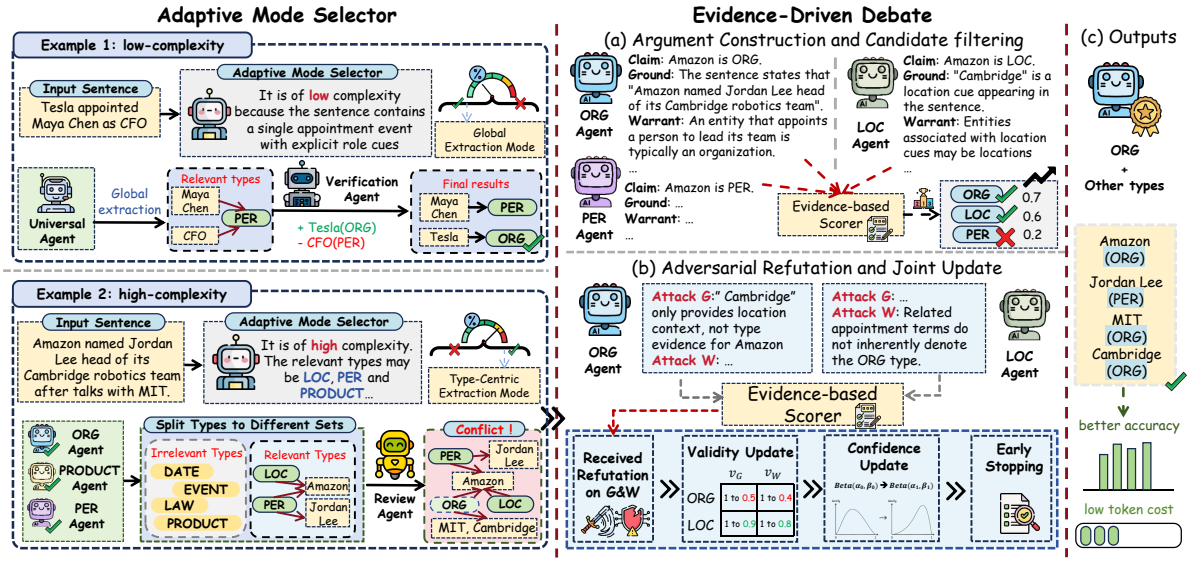

Figure 2: Overview of SMADE-IE. The Router Agent selects extraction modes. For conflicted predictions, Evidence-Driven Debate performs Toulmin-based ranking, Bayesian debate, and conflict resolution.

relation $r_{ij} \in \mathcal{T}_r$ that holds between them, producing a set of relations $R = \{(e_i, r_{ij}, e_j) \mid r_{ij} \in \mathcal{T}_r\}$. Finally, JERE simultaneously recognizes all entities and relations, yielding a set of fully typed quintuples $R^* = \{(e_i, t_e^i, r_{ij}, e_j, t_e^j)\}$, where each quintuple specifies a head entity, a tail entity, their types, and the relation connecting them. Since zero-shot NER and RE are handled by the same LLM framework in SMADE-IE and differ only in the task-definition prompt, we present SMADE-IE from the zero-shot NER perspective.

2.2 Adaptive Mode Selector

SMADE-IE dynamically selects the inference extraction mode for each input, reducing redundant calls while preserving type-level scrutiny in ambiguous cases. Inspired by the fast-and-slow thinking paradigm in LLM reasoning (Zhang et al., 2025a), we introduce a Router Agent to estimate both the relevant type subset and the sample complexity. Given a sentence s and the entity type set $\mathcal{T}_e$, the Router Agent produces,

$$(\mathcal{A}, c) = \mathrm{LLM}(s, P_{\mathcal{T}_e}^{sel}), \tag{1}$$

where $\mathrm{LLM}(\cdot)$ denotes the large language model, $P_{\mathcal{T}_e}^{sel}$ is the selector prompt, $\mathcal{A} \subseteq \mathcal{T}_e$ is the candidate type subset, and $c \in \{\mathrm{low}, \mathrm{med}, \mathrm{high}\}$ denotes the estimated sample complexity. We adopt three complexity levels instead of binary labels to better capture borderline uncertainty (Mahajan et al., 2026). The estimates are then mapped into two execution modes: low-complexity inputs use the lightweight Global Extraction Mode, while medium- and high-complexity inputs use the Type-Centric Extraction Mode.

2.3 Global Extraction Mode

To effectively balance extraction performance and computational cost, particularly in terms of token usage, SMADE-IE adopts the Global Extraction Mode for low-complexity inputs ($c = \mathrm{low}$). It employs two sequential LLM agents: a Universal Agent that first generates a one-shot candidate set, followed by a Verification Agent that performs targeted verification and refinement.

2.3.1 Universal Agent

For simple samples, we first apply the monolithic-prompt method (Xie et al., 2023) to extract all candidate entities with types in $\mathcal{T}_e$. Formally, given the global extraction prompt $P_{\mathcal{T}_e}^{\mathrm{global}}$, the output of the Universal Agent is defined as,

$$E_U = \mathrm{LLM}(s, P_{\mathcal{T}_e}^{\mathrm{global}}), \tag{2}$$

where every element $(e_i, t_e^i) \in E_U$ pairs a candidate entity with a type.

2.3.2 Verification Agent

Although the Universal Agent extracts all elements in a single pass, its coarse-grained generation strategy may still miss valid entities or introduce redundant candidates. To improve extraction reliability, the Verification Agent, inspired by the clean-up module in G&O (Li et al., 2024), performs bidirectional refinement over the candidate set E_U. Given the verification prompt $P_{\mathcal{T}_e}^{\mathrm{ver}}$ and E_U, the Verification Agent inserts missing entities and removes

unsupported candidates:

$$(E^+, E^-) = \mathrm{LLM}(s, E_U, P^{\mathrm{ver}}_{\mathcal{T}_e}), \tag{3}$$

where E^+ and E^- denote the sets of entities to be inserted and deleted, respectively. Accordingly, the final output of the Global Extraction Mode is:

$$E = (E_U \setminus E^-) \cup E^+. \tag{4}$$

2.4 Type-Centric Extraction Mode

For medium- and high-complexity inputs, SMADE-IE instantiates Type-Specific Agents for candidate types $\mathcal{A}$ selected by the Adaptive Mode Selector in Eq. (1) and adds a Review Agent to recover residual-type candidates. To resolve conflicts among agents, we introduce the Evidence-Driven Debate module that limits unconstrained discussion and weak-candidate interference, framing conflict resolution as evidence-based confidence comparison.

2.4.1 Type-Specific Agent

For each candidate type $t^i_{\mathcal{A}} \in \mathcal{A}$, we design a dedicated Type-Specific Agent with prompt $P^i_{\mathcal{A}}$. Given the input sentence s, the agent extracts candidate entities $e^i_{\mathcal{A}}$ associated with type $t^i_{\mathcal{A}}$:

$$e^i_{\mathcal{A}} \;=\; \mathrm{LLM}(s, P^i_{\mathcal{A}}), \tag{5}$$

where each element $e^{i,j}_{\mathcal{A}} \in e^i_{\mathcal{A}}$ denotes a candidate entity extracted for type $t^i_{\mathcal{A}}$. As illustrated in Figure 2, the PERSON agent extracts both **Amazon** and **Jordan Lee**. We then assign each extracted entity $e^{i,j}_{\mathcal{A}}$ to its corresponding type $t^i_{\mathcal{A}}$ and aggregate all predictions to obtain the final output of the Type-Specific Agents, denoted as $E_{\mathcal{A}}$.

2.4.2 Review Agent

The Adaptive Mode Selector estimates $\mathcal{A}$ in a single recall-oriented pass, but may miss peripheral types in long or context-rich sentences. To address this issue, we introduce the Review Agent over the residual type set $\mathcal{A}^* = \mathcal{T}_e \setminus \mathcal{A}$. Let $P^{\mathrm{rev}}_{\mathcal{A}^*}$ denote the Review Agent prompt restricted to $\mathcal{A}^*$. The Review Agent produces a residual candidate set:

$$E^*_{\mathcal{A}} = \mathrm{LLM}(s, P^{\mathrm{rev}}_{\mathcal{A}^*}), \tag{6}$$

The final output is defined as the union of the initial and residual sets:

$$\tilde{E}_{\mathcal{A}} = E_{\mathcal{A}} \cup E^*_{\mathcal{A}}. \tag{7}$$

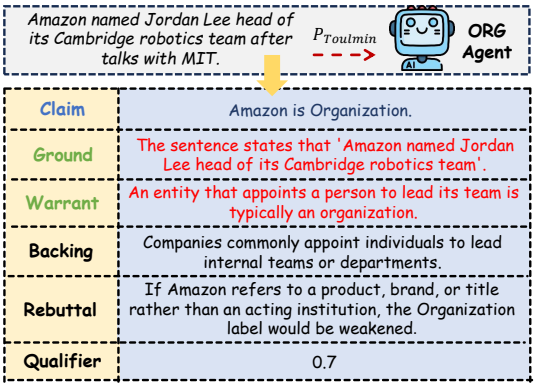

Figure 3: Example of a Toulmin-guided argument for the contested entity "Amazon", generated by the ORGANIZATION agent.

where each element in $\tilde{E}_{\mathcal{A}}$ is represented as a tuple $(\tilde{e}^i, \tilde{t}^i)$ containing a candidate entity and its assigned type. Since different Type-Specific Agents may assign different types to the same entity, we perform conflict detection before downstream processing. Specifically, if an entity $\tilde{e}^i$ is assigned multiple types, it is placed into the conflict set $\tilde{E}^{\mathrm{conf}}_{\mathcal{A}}$; otherwise, it is placed into the clean set $\tilde{E}^{\mathrm{clean}}_{\mathcal{A}}$:

$$\begin{aligned} \tilde{E}^{\mathrm{clean}}_{\mathcal{A}} &= \{(\tilde{e}, \tilde{t}) \in \tilde{E}_{\mathcal{A}} \;\mid\; |\{\tilde{t}' \mid (\tilde{e}, \tilde{t}') \in \tilde{E}_{\mathcal{A}}\}| = 1\}, \\ \tilde{E}^{\mathrm{conf}}_{\mathcal{A}} &= \left\{(\tilde{e}, \mathcal{Y}_{\tilde{e}}) \mid \mathcal{Y}_{\tilde{e}} = \{\tilde{t} \mid (\tilde{e}, \tilde{t}) \in \tilde{E}_{\mathcal{A}}\},\ |\mathcal{Y}_{\tilde{e}}| > 1\right\}. \end{aligned} \tag{8}$$

The clean set $\tilde{E}^{\mathrm{clean}}_{\mathcal{A}}$ is directly output, while the conflict set $\tilde{E}^{\mathrm{conf}}_{\mathcal{A}}$ is forwarded to the Evidence-Driven Debate module for conflict resolution and final type confirmation. Each element in $\tilde{E}^{\mathrm{conf}}_{\mathcal{A}}$ is represented as $(\tilde{e}^i, \{\tilde{t}^i_1, \ldots, \tilde{t}^i_N\})$, where $\tilde{e}^i$ denotes a candidate entity and $\{\tilde{t}^i_1, \ldots, \tilde{t}^i_N\}$ are its conflicting type assignments, and N is the number of conflicting types assigned to the entity.

2.4.3 Evidence-Driven Debate

This module enhances the robustness and evidence-groundedness of conflict adjudication by mitigating unsupported reasoning and low-confidence interference in free-form multi-agent debates. Specifically, each conflicting type assignment is recast as a Toulmin-style argument to enable structured evidence inspection. An external evidence scorer estimates candidate confidence from sentence-level support, ranks candidates for pre-debate filtering, and computes evidence-based attack scores during the debate. Support and refutation signals are then aggregated through Beta updates, and the debate terminates according to posterior stability and leading-candidate dominance.

Toulmin-Guided Argument Construction Recent work uses Toulmin theory to make implicit arguments explicit and improve the logical validity of LLM reasoning (Gupta et al., 2024; Xiao et al., 2024). Following this paradigm, we construct a structured argument for each conflicting candidate entity $\tilde{e}^i \in \tilde{E}_{\mathcal{A}}^{\text{conf}}$ and its associated types $\{\tilde{t}_1^i, \ldots, \tilde{t}_N^i\}$ to constrain LLM discourse. Each agent organizes its argument into five components: Claim (C), Ground (G), Warrant (W), Backing (B), and Rebuttal (R). For a conflicting entity $\tilde{e}^i$, the Toulmin-guided argument for conflict type $\tilde{t}_n^i$ is defined as:

$$(C_n, G_n, W_n, B_n, R_n) = \text{LLM}(s, \tilde{e}^i, \tilde{t}_n^i, P_{\text{Toulmin}}), \quad (9)$$

where C_n is the claim for $\tilde{t}_n^i$, G_n provides evidence from sentence s, W_n explains how G_n supports C_n, B_n gives auxiliary support, and R_n states a possible rebuttal. To measure how well s supports each argument, we compute a confidence score Q_n using an evidence scorer ϕ (Zha et al., 2023). Specifically, we concatenate $\{C_n, G_n, W_n, B_n\}$ into Para_n and construct a shared context $\hat{s}$ from sentence s and all conflicting type definitions $\{\tilde{t}_1^i, \ldots, \tilde{t}_N^i\}$,

$$Q_n = \phi(\hat{s}, \text{Para}_n) \in [0, 1], \quad (10)$$

where ϕ denotes an external evidence scorer (Zha et al., 2023) that measures how strongly the textual evidence (G_n, W_n, B_n) supports the claim C_n. Each agent is represented as the tuple $\mathcal{D}_n = \{C_n, G_n, W_n, B_n, R_n, Q_n\}$, which forms a structured, evidence-grounded argument. As shown in Figure 3, an ORGANIZATION agent may use this format to justify the contested entity **Amazon**. Considering all candidate types for a contested entity $\tilde{e}^i$ incurs $\mathcal{O}(N^2)$ pairwise attacks per debate round. To reduce computational complexity and mitigate noise from weak candidates, we rank the N Toulmin-guided arguments $\{\mathcal{D}_n\}_{n=1}^N$ by their qualifiers Q_n and retain only the top two:

$$\mathcal{D}_L = (\mathcal{D}_{\text{top1}}, \mathcal{D}_{\text{top2}}) = \text{Top2}_{\mathcal{D}_n}\left(\{\mathcal{D}_n\}_{n=1}^N;\ Q_n\right). \quad (11)$$

Bayesian Confidence Modeling We model candidate confidence with a Beta posterior $\text{Beta}(\alpha_l, \beta_l)$ for each agent $\mathcal{D}_l$ (Hu et al., 2025). The posterior accumulates supporting (α_l) and refuting (β_l) evidence across debate iterations, producing calibrated scores comparable across candidates. For candidate l, we compute an alignment-based scaling factor:

$$\rho_l = 1 - \phi(x_l, R_l),$$
$$\kappa_l = \kappa_{\min} + \rho_l(\kappa_{\max} - \kappa_{\min}), \quad (12)$$

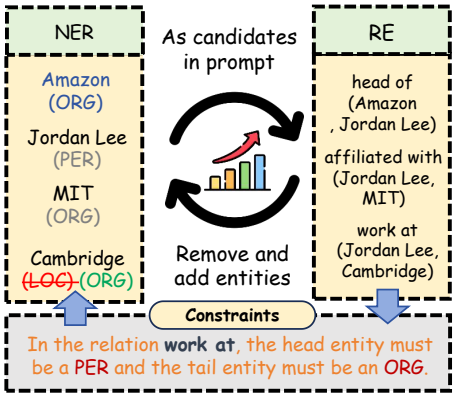

Figure 4: Iterative Entity–Relation Alignment under ontology constraints.

where x_l concatenates s with the candidate type definition $\tilde{t}_l^i$, and κ_l controls prior strength. Higher $\phi(x_l, R_l)$ implies weaker self-criticism and a smaller κ_l. The posterior is initialized with Q_l:

$$\alpha_l^{(0)} = Q_l \kappa_l, \qquad \beta_l^{(0)} = (1 - Q_l)\kappa_l, \quad (13)$$

The initialized posterior parameters are then updated through the joint evidence and confidence updating process described below.

Evidence-Grounded Adversarial Refutation To ensure evidence-grounded refutations, we restrict the attack surface to the two Toulmin components that directly support candidate l: Ground (G_l) and Warrant (W_l). Ground provides supporting evidence, while Warrant links the evidence to the claim C_l; other components offer contextual or auxiliary justification. Let $m \in \{G, W\}$ denote an attackable component, and let $z_{l,m}$ be its textual content for candidate l. Each component has a validity score $v_{l,m}^{(t)} \in [0, 1]$, initialized as $v_{l,m}^{(0)} = 1$, indicating its remaining credibility at iteration t. A component remains attackable while $v_{l,m}^{(t)} > 0$ and is removed once its validity reaches zero:

$$\mathcal{V}_l^{(t)} = \left\{ m \in \{G, W\} \;\middle|\; v_{l,m}^{(t)} > 0 \right\}. \quad (14)$$

At iteration t, the opposing agent $\mathcal{D}_{\bar{l}}$ attacks the defended agent $\mathcal{D}_l$ using sentence s, the active component set $\mathcal{V}_l^{(t)}$, and the directed attack prompt $P_{\bar{l}}^{\text{atk}}$. For each defended component $m \in \mathcal{V}_l^{(t)}$, it generates a refutation $u_{\bar{l},m}$:

$$u_{\bar{l},m} = \text{LLM}\left(s,\ \mathcal{V}_l^{(t)},\ P_{\bar{l},m}^{\text{atk}}\right). \quad (15)$$

We measure attack effectiveness by how much the refutation weakens the defended component

$z_{l,m}$ under the type-specific context x_l. The attack strength at iteration t is

$$a_{l,m}^{(t)} = \sigma\left(\gamma\left[\phi(x_l,\, u_{\bar{l},m}) - \phi(x_l,\, z_{l,m})\right]\right), \quad (16)$$

where γ is a temperature parameter and $\sigma(\cdot)$ is the sigmoid function. Larger $a_{l,m}^{(t)}$ indicates stronger support for the refutation relative to the defended content.

Joint Evidence and Confidence Updating To prevent posterior recovery through reformulated invalid components, we couple component validity with the Beta posterior via the shared attack score $a_{l,m}^{(t)}$. This ties posterior updates to evidence credibility and keeps both update processes consistent. For each component, $v_{l,m}^{(t)}$ is updated in two stages. First, the attack induces exponential decay:

$$\tilde{v}_{l,m}^{(t+1)} = v_{l,m}^{(t)} \exp\Big(-\eta\, a_{l,m}^{(t)}\Big), \quad (17)$$

where $\eta > 0$ is the decay rate. Given a validity activity threshold θ_v, the realized validity is

$$v_{l,m}^{(t+1)} = \begin{cases} \tilde{v}_{l,m}^{(t+1)}, & \tilde{v}_{l,m}^{(t+1)} > \theta_v, \\ \dfrac{\theta_v+1}{2}\,\omega^{k_m^l}, & \begin{array}{r} 0 < \tilde{v}_{l,m}^{(t+1)} \le \theta_v, \\ \text{and } \dfrac{\theta_v+1}{2}\,\omega^{k_m^l} \ge \theta_v, \end{array} \\ 0, & \text{otherwise.} \end{cases} \quad (18)$$

where $\tilde{v}_{l,m}^{(t+1)}$ is retained if it remains above θ_v. Otherwise, the component enters revision and is reset to the discounted ceiling $\frac{\theta_v+1}{2}\omega^{k_m^l}$, where k_m^l is the number of prior revisions and $\omega \in (0,1)$ is the revision discount factor. Repeated revisions therefore reduce credibility. If the revised validity still exceeds θ_v, the component remains in $\mathcal{V}_l$; otherwise, it is permanently removed from the attack surface. The same attack score updates the Beta posterior of candidate l by adding refutation mass to β_l and residual support mass to α_l:

$$\begin{aligned} \beta_l^{(t+1)} &= \beta_l^{(t)} + \frac{\bar{v}_{\bar{l}}^{(t)}}{\left|\mathcal{V}_l^{(t)}\right|} \sum_{m\in\mathcal{V}_l^{(t)}} a_{l,m}^{(t)} \cdot v_{l,m}^{(t)} \\ \alpha_l^{(t+1)} &= \alpha_l^{(t)} + \frac{1}{\left|\mathcal{V}_l^{(t)}\right|} \sum_{m\in\mathcal{V}_l^{(t)}} \big(1 - a_{l,m}^{(t)}\big) \cdot v_{l,m}^{(t)} \end{aligned} \quad (19)$$

where $\bar{v}_{\bar{l}}^{(t)}$ denotes the average validity of the opponent's active evidence components:

$$\bar{v}_{\bar{l}}^{(t)} = \frac{1}{|\mathcal{V}_{\bar{l}}^{(t)}|} \sum_{m\in\mathcal{V}_{\bar{l}}^{(t)}} v_{\bar{l},m}^{(t)}. \quad (20)$$

The weights have distinct roles: $\bar{v}_{\bar{l}}^{(t)}$ downweights attacks from opponents with weakened evidence, while $v_{l,m}^{(t)}$ reduces the influence of components weakened in earlier rounds.

Dual-Track Early Stopping To avoid redundant debate rounds, we terminate the process when either **posterior stability** or **leading candidate confidence** is satisfied. Both criteria operate on the Beta posteriors induced by the joint update rule:

$$\pi_l^{(t)} \sim \mathrm{Beta}\Big(\alpha_l^{(t)}, \beta_l^{(t)}\Big), \quad (21)$$

for each candidate $l \in L$ at iteration t.

The **posterior stability** criterion measures convergence between consecutive posterior distributions using the average squared Hellinger distance:

$$\Delta_{\mathrm{conf}}^{(t)} = \frac{1}{2}\sum_{l\in L} H^2\Big(\pi_l^{(t)}, \pi_l^{(t-1)}\Big) < \varepsilon, \quad (22)$$

where $H^2(\cdot)$ denotes squared Hellinger distance. The debate terminates once $\Delta_{\mathrm{conf}}^{(t)} < \varepsilon$, indicating that the posteriors have effectively converged.

The **leading candidate confidence** criterion stops the debate when the leading candidate is sufficiently likely to outperform its opponent. For scalar comparison, we use the posterior mean

$$\hat{p}_l^{(t)} = \frac{\alpha_l^{(t)}}{\alpha_l^{(t)} + \beta_l^{(t)}}, \quad (23)$$

which integrates all supporting and refuting evidence accumulated up to iteration t. Let

$$l^* = \arg\max_{l\in L} \hat{p}_l^{(t)}, \quad (24)$$

where l^* denotes the leading candidate at iteration t, and $\bar{l}^*$ its opponent. The debate terminates if

$$\mathcal{P}\Big(\pi_{l^*}^{(t)} > \pi_{\bar{l}^*}^{(t)}\Big) > \delta_{\mathrm{stop}}. \quad (25)$$

If neither criterion is satisfied within $T_{\max}$ rounds, the final winner is selected using the scorer ϕ:

$$l^{\mathrm{winner}} = \arg\max_{l\in L} \phi(\hat{s}, \mathrm{Para}_l). \quad (26)$$

Appendix D provides a Chebyshev-style bound motivating the dual-track stopping rule.

2.5 Iterative Entity–Relation Alignment

For JERE, we enforce ontology-consistent alignment between E and R, as shown in Figure 4. At iteration t, the procedure proceeds as follows.

Method	CoNLL03 (3)		OntoNotes5 (18)		SciERC (5)		CrossRE (38)		REDFM (10)		Avg.	
	$F1_P$	$F1_S$	$F1_P$	$F1_S$	$F1_P$	$F1_S$	$F1_P$	$F1_S$	$F1_P$	$F1_S$	$F1_P$	$F1_S$
AEiO ♢	65.24	61.17	56.88	43.12	35.36	23.80	34.61	30.80	21.14	17.45	42.65	35.27
One-Step ♣	77.29	73.02	32.38	25.72	27.89	21.77	25.45	23.65	25.21	18.93	37.64	32.62
G&O ♣	76.36	72.35	31.65	23.40	31.83	24.05	27.00	25.38	27.07	20.61	38.78	33.16
CROSSAGENTIE♠	73.15	70.93	45.30	37.18	35.96	25.36	46.15	44.40	29.20	25.16	45.95	40.61
SMADE-IE♠(Ours)	**77.33**	**74.72**	**66.08**	**52.05**	**46.56**	**37.80**	**65.50**	**60.56**	**31.12**	**27.13**	**57.32**	**50.45**

Table 2: NER results on five benchmark datasets. Parentheses indicate the number of entity types (e.g., CoNLL03 has 3 entity types). **Bold** and underlined values denote the best and second-best results, respectively. ♢, ♣, and ♠ indicate monolithic prompting, each-type prompting, and multi-agent debate methods, respectively.

Method	DocRED (32)		SemEval2010 (9)		SciERC (8)		CrossRE (17)		REDFM (32)		Avg.	
	$F1_P$	$F1_S$	$F1_P$	$F1_S$	$F1_P$	$F1_S$	$F1_P$	$F1_S$	$F1_P$	$F1_S$	$F1_P$	$F1_S$
AEiO ♢	30.32	24.84	27.55	9.86	19.98	10.91	4.79	3.56	10.04	6.24	18.54	11.08
One-Step ♣	22.15	18.41	16.13	5.16	18.56	8.94	12.00	8.31	7.11	5.34	15.19	9.23
G&O ♣	19.69	16.71	12.16	4.56	13.07	5.58	8.90	7.01	5.65	4.12	11.89	7.60
CROSSAGENTIE ♠	13.72	11.49	11.77	2.91	14.61	8.79	11.81	8.57	5.36	3.44	11.45	7.04
SMADE-IE ♠	**30.78**	**26.99**	**29.54**	**16.76**	**21.47**	**11.20**	**16.22**	**12.25**	**13.83**	**11.34**	**22.37**	**15.71**

Table 3: RE results. SMADE-IE obtains the strongest macro average across all relation-extraction benchmarks

Method	CoNLL04 (5)		NYT (7)		Avg.	
	$F1_P$	$F1_S$	$F1_P$	$F1_S$	$F1_P$	$F1_S$
CROSSAGENTIE ♠	42.50	29.05	17.23	16.77	29.87	22.91
SMADE-IE ♠	**58.44**	**44.74**	**30.22**	**29.22**	**44.33**	**36.98**

Table 4: JERE results (%); $F1_P$ / $F1_S$ as in Table 2. Both types and the relation of head and tail type must match.

Entity completion Given the candidate relation set $\bar{R}^{(t)} = RE(E^{(t)})$, we add entities that appear in $\bar{R}^{(t)}$ but are missing from $E^{(t)}$ after inferring their types:

$$\bar{E}^{(t+1)} = E^{(t)} \cup \{\, (e,t) \notin E^{(t)} \mid e \in \xi(\bar{R}^{(t)}) \,\}, \quad (27)$$

where $\xi(\cdot)$ denotes the entities involved in the candidate relations. This step completes missing entities under relational evidence.

Consistency correction We further enforce ontology consistency on $\bar{R}^{(t)}$ with respect to $\bar{E}^{(t+1)}$. For each violation, the Consistency Agent either revises the conflicting entity type if the relation type is trusted, or blacklists the relation otherwise. The aligned relations and entities are updated as:

$$R^{(t)} = \bar{R}^{(t)} \setminus B^{(t)}. \quad (28)$$

Entities not involved in any remaining relation are removed:

$$E^{(t+1)} = \{\, (e,t) \in \bar{E}^{(t+1)} \mid e \in \xi(R^{(t)}) \,\}. \quad (29)$$

We regenerate candidate relations from the updated entity set and filter them using the updated blacklist:

$$R^{(t+1)} = \bar{R}^{(t+1)} \setminus B^{(t+1)}. \quad (30)$$

The procedure terminates once all violations are resolved or K_{JERE} is reached, yielding the aligned JERE output.

3 Experiment and Analysis

This section presents the main experimental results and analyses. Additional details on datasets, evaluation metrics, baselines, implementation settings, hyperparameters, and prompts are provided in Appendices B–F.

3.1 Main Results

Tables 2 and 3 report NER and RE results on five benchmarks using GPT-3.5-Turbo-0125 as the backbone. SMADE-IE achieves the best performance across both tasks, demonstrating the effectiveness of multi-agent deliberation for information extraction. For NER, SMADE-IE outperforms all baselines, with average $F1_P$ gains of 14.67 over AEiO, 19.68 over One-Step, 18.54 over G&O, and 11.37 over CROSSAGENTIE. The gains are larger on datasets with richer type inventories, surpassing CROSSAGENTIE by 20.78 $F1_P$ on OntoNotes5, 19.35 on CrossRE, and 10.60 on SciERC. For RE, SMADE-IE obtains the highest $F1_P$ and $F1_S$ scores on all benchmarks, improving average $F1_P$ by 3.83, 7.18, 10.48, and 10.92 over AEiO, One-Step, G&O, and CROSSAGENTIE, respectively. Its largest gain over CROSSAGENTIE occurs on SemEval2010 (+17.77 $F1_P$), while consistent improvements on DocRED, SciERC, CrossRE, and

Variant	CoNLL04		NYT	
	$F1_P$	$F1_S$	$F1_P$	$F1_S$
Full SMADE-IE	58.44	**44.74**	**30.22**	**29.22**
w/o *IERA*	59.03	42.73	24.94	22.47
Type-Centric Only	**59.43**	40.33	29.65	28.47
w/o *Relevant Type Selection*	48.66	34.18	27.43	26.43
w/o *Review Agent*	52.75	35.12	29.50	27.56
w/o *Debate*	54.13	38.02	24.24	23.38
Global Only	52.47	40.20	16.28	14.27
w/o *Verification Agent*	46.92	35.19	12.02	11.16

Table 5: Ablation results on JERE. The first block evaluates *IERA*, while the remaining blocks analyze Adaptive Mode Selection, Verification Agent, Review Agent, and Evidence-Driven Debate.

REDFM indicate strong generalizability across relation extraction settings.

For the JERE task, which requires joint entity-relation extraction and cross-task completion, we primarily compare with CROSSAGENTIE. As shown in Table 4, SMADE-IE improves over CROSSAGENTIE by 14.46 $F1_P$ and 14.07 $F1_S$ on average. The gains are consistent on both datasets, with $F1_P$ increasing from 42.50 to 58.44 on CoNLL04 and from 17.23 to 30.22 on NYT.

3.2 Ablation Studies

Table 5 presents ablation results on CoNLL04 and NYT, grouped by module functionality. Removing *Iterative Entity–Relation Alignment* (*IERA*) substantially degrades performance on NYT ($F1_P$ 30.22→24.94, $F1_S$ 29.22→22.47), with a smaller effect on CoNLL04, indicating that IERA is beneficial when entity–relation inconsistencies are frequent. Comparisons with the monolithic-prompt baseline further validate the effectiveness of *Adaptive Mode Selection*. It consistently improves $F1_S$ across datasets, outperforming *Global Extraction Mode* by 4.54 on CoNLL04 and 14.95 on NYT, and *Type-Centric Only* by 4.41 and 0.75, respectively.

Within each mode, ablations highlight the importance of internal components: in *Global Extraction Mode*, the *Verification Agent* boosts precision ($F1_P$ drops from 52.47 → 46.92 on CoNLL04 when removed); in *Type-Centric Mode*, *Relevant Type Selection*, *Review Agent*, and *Evidence-Driven Debate* (*Debate*) all contribute, with the latter particularly critical on NYT ($F1_P$ 29.65 → 24.24, $F1_S$ 28.47 → 23.38) and the former on CoNLL04 ($F1_P$ 59.43 → 48.66). These results confirm that each module plays a complementary role in improving precision and structured F1 across datasets.

Method	NER		RE	
	CoNLL03 (3)	**REDFM** (10)	**DocRED** (32)	**CrossRE** (17)
AEiO	1078	2163	5460	4798
One-Step	1407	7673	7615	6846
G&O	2304	14202	9271	14312
CrossAgentIE	1508	13480	21784	18923
SMADE-IE	1094	11514	3240	3765

Table 6: Average total token cost per sample across different datasets.

3.3 Token Efficiency

Table 6 reports the average total token cost per sample on selected NER and RE datasets. On CoNLL03, SMADE-IE uses 1,094 tokens, close to AEiO (1,078) and lower than One-Step (1,407), G&O (2,304), and CROSSAGENTIE (1,508). On REDFM NER, SMADE-IE also reduces cost compared with G&O and CROSSAGENTIE, using 11,514 tokens versus 14,202 and 13,480. The savings are more pronounced on RE: SMADE-IE uses 3,240 tokens on DocRED and 3,765 on CrossRE, far below CROSSAGENTIE's 21,784 and 18,923 tokens. It is also more efficient than One-Step and G&O on both RE datasets. These results show that sparse type selection avoids unnecessary full-schema agent calls, keeping the cost close to lightweight prompting on simple cases while substantially reducing multi-agent debate overhead on relation-heavy datasets.

4 Conclusion

This paper proposes SMADE-IE, a sparse and evidence-driven multi-agent framework for zero-shot information extraction. SMADE-IE employs the Adaptive Mode Selector to dynamically route inputs into the lightweight Global Extraction Mode or the Type-Centric Extraction Mode, reducing redundant agent interactions and reasoning noise. For conflicting predictions, we further introduce the Evidence-Driven Debate mechanism with Toulmin-style argument modeling and Bayesian confidence updates. In addition, the Iterative Entity–Relation Alignment strategy is designed for joint extraction tasks. Experimental results on 9 benchmark datasets across NER, RE, and JERE tasks demonstrate that SMADE-IE consistently outperforms existing zero-shot IE baselines while improving token efficiency. Future work will explore more efficient debate scheduling and broader structured extraction settings.

Limitations

SMADE-IE has three main limitations. First, the Evidence-Driven Debate relies on a frozen external evidence scorer. Its NLI calibration upper-bounds the reliability of Bayesian confidence updates. Second, although the Adaptive Mode Selector keeps token cost close to monolithic-prompt baselines on simple inputs, the Type-Centric Extraction Mode still issues multiple LLM calls per cross-type conflict, and its efficiency advantage diminishes on long-document, type-dense datasets such as REDFM, where invoked debates are only halved rather than reduced by an order of magnitude. Third, our main experiments are conducted with GPT-3.5-Turbo-0125 and cross-backbone evidence is restricted to gemini-3-flash-preview on a subset of datasets, leaving smaller open-source models, longer documents, and multilingual schemas as natural directions for future work.

Ethics Statement

In this work, we propose SMADE-IE, a sparse multi-agent framework designed to improve large language model performance on fundamental information extraction tasks, including named entity recognition and relation extraction. All our analyses and evaluations are based on standard, publicly available academic datasets. We do not anticipate any specific ethical issues or negative societal impacts regarding the methodology or the topics of this research.

References

Elisa Bassignana and Barbara Plank. 2022. CrossRE: A cross-domain dataset for relation extraction. In *Findings of the EMNLP 2022*, pages 3592–3604.

Zhijun Chen, Hailong Sun, Wanhao Zhang, Chunyi Xu, Qianren Mao, and Pengpeng Chen. 2023. Neural-hidden-crf: A robust weakly-supervised sequence labeler. In *Proceedings of the ACM SIGKDD 2023*, page 274–285. ACM.

Hyeong Kyu Choi, Jerry Zhu, and Sharon Li. 2026. Debate or vote: Which yields better decisions in multi-agent large language models? In *Proceedings of the NIPS 2026*.

Wei Fan, JinYi Yoon, and Bo Ji. 2026. imad: Intelligent multi-agent debate for efficient and accurate llm inference. *Proceedings of the AAAI 2026*, 40(35):29403–29411.

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

Chufan Gao, Xulin Fan, Jimeng Sun, and Xuan Wang. 2024. PromptRE: Weakly-supervised document-level relation extraction via prompting-based data programming. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 132–145.

Yong Guan, Hao Peng, Lei Hou, and Juanzi Li. 2025. MMD-ERE: Multi-agent multi-sided debate for event relation extraction. In *Proceedings of the COLING 2025*, pages 6889–6896.

Ankita Gupta, Ethan Zuckerman, and Brendan O'Connor. 2024. Harnessing toulmin's theory for zero-shot argument explication. In *Proceedings of the ACL 2024*, pages 10259–10276.

Ridong Han, Chaohao Yang, Tao Peng, Prayag Tiwari, Xiang Wan, Lu Liu, and Benyou Wang. 2024. An empirical study on information extraction using large language models. *Preprint*, arXiv:2409.00369.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38.

Tianyu Hu, Zhen Tan, Song Wang, Huaizhi Qu, and Tianlong Chen. 2025. Multi-agent debate for llm judges with adaptive stability detection. In *Advances in Neural Information Processing Systems*, pages 46504–46540.

Pere-Lluís Huguet Cabot, Simone Tedeschi, Axel-Cyrille Ngonga Ngomo, and Roberto Navigli. 2023. RED^{fm}: A filtered and multilingual relation extraction dataset. In *Proceedings of the ACL 2023*, pages 4326–4343.

Nitisha Jain. 2020. Domain-specific knowledge graph construction for semantic analysis. In *Proceedings of the Semantic Web: ESWC 2020*, pages 250–260.

Nanda Kambhatla. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL) Companion Volume*, pages 22–25.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*, pages 282–289.

Guozheng Li, Peng Wang, and Wenjun Ke. 2023a. Revisiting large language models as zero-shot relation extractors. In *Findings of the EMNLP 2023*, pages 6877–6892.

Yinghao Li, Colin Lockard, Prashant Shiralkar, and Chao Zhang. 2023b. Extracting shopping interest-related product types from the web. In *Findings of the ACL 2023*, pages 7509–7525.

Yinghao Li, Rampi Ramprasad, and Chao Zhang. 2024. A simple but effective approach to improve structured language model output for information extraction. In *Findings of the EMNLP 2024*, pages 5133–5148.

Yinghao Li, Le Song, and Chao Zhang. 2022. Sparse conditional hidden markov model for weakly supervised named entity recognition. In *Proceedings of the ACM SIGKDD 2022*, page 978–988.

Pierre Lison, Jeremy Barnes, and Aliaksandr Hubin. 2021. skweak: Weak supervision made easy for nlp. In *Proceedings of the ACL 2021*, page 337–346.

Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu. 2020. Event extraction as machine reading comprehension. In *Proceedings of the EMNLP 2020*, pages 1641–1651.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Jie Lou, Yaojie Lu, Dai Dai, Wei Jia, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2023. Universal information extraction as unified semantic matching. In *Proceedings of the AAAI 2023*, pages 13318–13326.

Meng Lu, Yuzhang Xie, Zhenyu Bi, Shuxiang Cao, and Xuan Wang. 2025. CrossAgentIE: Cross-type and cross-task multi-agent LLM collaboration for zero-shot information extraction. In *Findings of the ACL 2025*, pages 13953–13977.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified structure generation for universal information extraction. In *Proceedings of the ACL 2022*, pages 5755–5772.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the EMNLP 2018*, pages 3219–3232.

Yujie Luo, Xiangyuan Ru, Kangwei Liu, Lin Yuan, Mengshu Sun, Ningyu Zhang, Lei Liang, Zhiqiang Zhang, Jun Zhou, Lanning Wei, and 1 others. 2025. Oneke: A dockerized schema-guided llm agent-based knowledge extraction system. In *Companion Proceedings of the ACM WWW 2025*, pages 2871–2874.

Pranav Mahajan, Ihor Kendiukhov, Syed Hussain, and Lydia Nottingham. 2026. Mind the gap: How elicitation protocols shape the stated-revealed preference gap in language models. *Preprint*, arXiv:2601.21975.

Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the Seventh Conference on Natural Language Learning (CoNLL)*, pages 188–191.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the CoNLL 2013*, pages 143–152.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Proceedings of the ECML-PKDD 2010*, pages 148–163.

Bryan Rink and Sanda Harabagiu. 2010. UTD: Classifying semantic relations by combining machine learning and rich linguistic features. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval)*, pages 256–259.

Dan Roth and Wen-tau Yih. 2004. A linear programming formulation for global inference in natural language tasks. In *Proceedings of the NAACL 2004*, pages 1–8.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lacalle, German Rigau, and Eneko Agirre. 2024. Gollie: Annotation guidelines improve zero-shot information-extraction. In *Proceedings of the ICLR 2024*, pages 47083–47107.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the ICML 2023*, pages 31210–31227.

Yuchen Shi, Guochao Jiang, Tian Qiu, and Deqing Yang. 2024. Agentre: An agent-based framework for navigating complex information landscapes in relation extraction. In *Proceedings of the ACM CIKM 2024*, page 2045–2055.

Hongduan Tian, Xiao Feng, Rolan Yan, Bo Han, and 1 others. 2026. Multi-agent debate with memory masking. In *The Fourteenth International Conference on Learning Representations*.

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 NAACL 2003*, pages 142–147.

Sijia Wang and Lifu Huang. 2024. Debate as optimization: Adaptive conformal prediction and diverse retrieval for event extraction. In *Findings of the EMNLP 2024*, pages 16422–16435.

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

Ruiyu Xiao, Lei Wu, Yuhang Gou, Weinan Zhang, and Ting Liu. 2024. Prove your point!: Bringing proof-enhancement principles to argumentative essay generation. In *EMNLP 2024*, pages 18995–19008.

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 EMNLP 2023*, pages 7935–7956.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. 2024. Large language models for generative information extraction: A survey. *Frontiers of Computer Science*, 18(6):186357.

Minglai Yang, Ethan Huang, Liang Zhang, Mihai Surdeanu, William Yang Wang, and Liangming Pan. 2025. How is llm reasoning distracted by irrelevant context? an analysis using a controlled benchmark. In *Proceedings of the EMNLP 2025*, pages 13340–13358.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the ACL 2019*, pages 764–777.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of the COLING 2014*, pages 2335–2344.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the ACL 2023*, pages 11328–11348.

Duzhen Zhang, Zhong-Zhi Li, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Xiuyi Chen, Yingying Zhang, and 1 others. 2025a. From system 1 to system 2: A survey of reasoning large language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zikang Zhang, Wangjie You, Tianci Wu, Xinrui Wang, Juntao Li, and Min Zhang. 2025b. A survey of generative information extraction. In *Proceedings of the COLING 2025*, pages 4840–4870.

Suncong Zheng, Feng Wang, Hongyun Bao, Yuexing Hao, Peng Zhou, and Bo Xu. 2017. Joint extraction of entities and relations based on a novel tagging scheme. In *Proceedings of the ACL 2017*, pages 1227–1236.

A Related Works

A.1 Supervision-based Information Extraction

Early IE methods mainly rely on supervised learning, typically modeling NER as sequence labeling and RE as classification or structured prediction (Lafferty et al., 2001; McCallum and Li, 2003; Kambhatla, 2004; Rink and Harabagiu, 2010). Neural IE methods reduce reliance on manual feature engineering by learning distributed representations and contextual encoders for sequence labeling, relation classification, and joint extraction (Zeng et al., 2014; Zheng et al., 2017). More recently, pretrained-language-model fine-tuning and instruction-tuning methods have further improved IE performance under supervised and transfer settings (Lu et al., 2022; Lou et al., 2023). However, these methods usually assume sufficient annotated data, stable domains, and fixed entity and relation schemas, making them difficult to adapt to annotation-scarce settings or dynamically changing schemas. Weakly and distantly supervised methods reduce annotation cost by using unlabeled corpora, external knowledge bases, or heuristic labeling functions (Li et al., 2022; Chen et al., 2023; Gao et al., 2024), but they still require additional supervision signals or manual design efforts and may suffer from noisy pseudo-labels (Lison et al., 2021). In contrast, this work focuses on zero-shot IE, where models perform extraction using only input text, type definitions, and output constraints. While zero-shot IE poses greater challenges than fully or semi-supervised methods, it offers superior adaptability to a dynamically changing world without requiring any additional fine-tuning.

A.2 LLM-based Zero-shot Information Extraction

Recent studies have explored the use of LLMs for zero-shot IE by prompting models with task instructions, type definitions, or output schemas (Xu et al., 2024; Zhang et al., 2025b). *Monolithic-prompt* methods extract all elements in a single pass, offering high efficiency but often suffering from type confusion, boundary errors, and missed mentions as the schema size increases (Xie et al., 2023; Li et al., 2024; Han et al., 2024). In contrast, *each-type prompt* methods query entity or relation types separately, enabling more focused extraction but frequently producing overlapping or conflicting outputs across prompts (Li et al., 2023a, 2024).

To address such cross-type inconsistencies, *multi-agent debate* methods instantiate type-specific agents and resolve conflicting predictions through inter-agent deliberation. Related research has further explored multi-agent collaboration, tool use, knowledge augmentation, and workflow orchestration for IE (Wang and Huang, 2024; Shi et al., 2024; Wang et al., 2025; Luo et al., 2025). For example, CROSSAGENTIE (Lu et al., 2025) assigns dedicated agents to entity and relation types and resolves structured prediction conflicts through cross-type and cross-task debate. However, these approaches select the full type schema for every input, causing token cost and contextual noise to scale with schema size. Moreover, their free-form debate mechanisms lack explicit modeling of disputed claims, evidence constraints, and confidence updates, which can lead to unreliable adjudication. In contrast, SMADE-IE selects only Router-selected types and replaces free-form deliberation with a Toulmin-structured, evidence-scored Bayesian debate framework.

B Experimental Setup

B.1 Datasets and Evaluation Metrics

We evaluate SMADE-IE on 12 zero-shot IE task settings over 9 benchmark datasets, spanning NER, RE, and JERE across general-domain, scientific, and document-level text. For NER, we use CoNLL03 (Tjong Kim Sang and De Meulder, 2003), OntoNotes5 (Pradhan et al., 2013), SciERC (Luan et al., 2018), CrossRE (Bassignana and Plank, 2022), and REDFM (Huguet Cabot et al., 2023). For RE, we evaluate on DocRED (Yao et al., 2019), SemEval-2010 Task 8 (Hendrickx et al., 2010), and reuse SciERC, CrossRE, and REDFM. For JERE, we use CoNLL04 (Roth and Yih, 2004) and NYT (Riedel et al., 2010). During inference, the model is provided only with natural-language ontology definitions as external knowledge, following the zero-shot evaluation protocol of Li et al. (2024) and Lu et al. (2025).

For datasets whose ontologies contain rare, ambiguous, or under-specified labels, we follow the evaluation settings adopted by prior work (Li et al., 2024). Specifically, CoNLL03 and CoNLL04 retain only the three concrete entity types. CrossRE retains 38 out of 39 entity types after removing the OTHER category and uses all 17 semantic relation types. REDFM and SciERC exclude only reserved generic labels. DocRED is restricted to

Statistic	CoNLL03	OntoNotes5	CoNLL04	NYT	CrossRE	REDFM	SciERC
# Selected Instances	300	200	288	369	200	446	269
Avg. Sequence Length	13	18	29	36	37	76	23
Avg. Entities per Instance	1.50	1.31	2.38	2.00	4.72	3.57	2.19
Entity Types	3	18	3	3	38	10	5

Table 7: Statistics of the NER benchmarks used in our evaluation. "# Selected Instances" denotes the number of evaluation samples used in our experiments, "Avg. Sequence Length" indicates the average number of tokens per instance, "Avg. Entities per Instance" refers to the average number of entity mentions in each sample and "Entity Types" refers to the number of entity categories evaluated in each dataset.

Statistic	DocRED	SemEval2010	CoNLL04	NYT	CrossRE	REDFM	SciERC
# Selected Instances	200	200	288	369	200	446	269
Avg. Sequence Length	30	17	29	36	37	76	23
Avg. Relations per Instance	2.46	0.84	1.41	0.72	5.05	2.77	1.88
Relation Types	32	9	5	7	17	32	8

Table 8: Statistics of the relation extraction (RE) benchmarks used in our evaluation. "# Selected Instances" denotes the number of evaluation samples used in our experiments, "Avg. Sequence Length" indicates the average number of tokens per instance, "Avg. Relations per Instance" refers to the average number of relation mentions in each sample, and "Relation Types" is reported as selected relation types / total ontology relation types when only a subset of relations is evaluated. All evaluated relations are binary.

the 32 most frequent relation types, while SemEval is limited to the 9 forward relation types. Detailed dataset entity-level statistics shown in Table 7 and relation-level statistics in Table 8. For datasets containing more than a few hundred test instances, we randomly sample a fixed evaluation subset shared across all baselines and ablation settings to ensure fair and directly comparable evaluations of token cost and F1 performance.

We report micro Partial F1 ($\mathrm{F1_P}$) and Strict F1 ($\mathrm{F1_S}$), together with averages across datasets. Partial F1 allows boundary overlap between predicted and gold spans, whereas Strict F1 requires exact span matching.

B.2 Baselines

We compare SMADE-IE against four representative zero-shot LLM-based IE methods spanning three paradigms. For *monolithic prompting*, **All-Entity-in-One (AEiO)** (Li et al., 2024) extracts all candidate types in a single prompt. For *each-type prompting*, **One-Step** (Li et al., 2024) issues a dedicated prompt for each entity or relation type, while **Generation and Organization (G&O)** (Li et al., 2024) decouples content generation from output organization. For *multi-agent debate*, **CrossAgentIE** (Lu et al., 2025) instantiates type-specific agents and resolves conflicts through cross-type and cross-task debate.

B.3 Implementation Details

All LLM-based methods use GPT-3.5-Turbo-0125 in a zero-shot setting, receiving only the input text, task schema, and type definitions. Agent interactions are implemented with AutoGen[1], and the external evidence scorer ϕ uses the AlignScore-base checkpoint (Zha et al., 2023)[2] deployed locally. For SMADE-IE, the maximum number of debate rounds is three, with $\delta_{\text{stop}} = 0.75$, $\varepsilon = 0.02$, and $\gamma = 8$. All hyperparameters are fixed across the twelve evaluation settings; the complete configuration is reported in Table 9.

For baselines, we use the official implementations of AEiO, One-Step, and G&O. We extend AEiO to RE by retaining its all-in-one chat structure and replacing only the output schema: a single GPT-3.5-Turbo-0125 session jointly extracts relation mentions, prunes ungrounded ones, and emits per-type Markdown tables with head, tail, and validity columns. Outputs are parsed by normalizing relation-type headers, discarding rows whose validity token is not "yes" or whose head and tail coincide, and mapping the rest to token spans via the spaCy and locate-phrase routines from the G&O codebase.

CrossAgentIE (Lu et al., 2025) has no public code and is re-implemented on AutoGen with the same backbone. Each type is bound to a Type-

[1] https://microsoft.github.io/autogen/
[2] https://huggingface.co/yzha/AlignScore

Symbol / Name	Value
$T_{\max}$ (max debate rounds)	3
$[\kappa_{\min}, \kappa_{\max}]$ (prior strength)	[0.8, 1.5]
θ_v (validity activity threshold)	0.6
η (validity decay coefficient)	0.8
ω (revision discount)	0.8
δ_{stop} (posterior superiority)	0.75
ε (confidence convergence)	0.02
γ (sigmoid temperature)	8.0
Non-Debate Agent temperature	0.1
Type-Specific Agent temperature	0.9
K_{JERE} (IERA iterations)	1

Table 9: Hyperparameters of SMADE-IE, held fixed across the twelve evaluation settings. Symbol names follow §2.4.3 and §2.5.

Method	OntoNotes5 (NER)		SemEval2010 (RE)		CoNLL04 (JERE)	
	$F1_P$	$F1_S$	$F1_P$	$F1_S$	$F1_P$	$F1_S$
AEiO ♢	79.37	68.76	**37.31**	21.64	–	–
One-Step ♣	27.23	20.96	15.39	4.31	–	–
G&O♣	24.47	19.14	12.65	3.91	–	–
CrossAgentIE♠	71.95	60.11	15.84	4.52	54.61	38.85
SMADE-IE♠	**88.14**	**72.95**	30.60	**23.73**	**62.09**	**43.69**

Table 10: Backbone generalization on gemini-3-flash-preview. "–" indicates that the corresponding method is not supported by this task.

Specific Agent that emits per-mention extractions with confidence scores via the original `### head ###/@@@ tail @@@` markers. When a head–tail pair is claimed by more than one type, the top-two claimants by confidence enter a `GroupChat` debate of up to two rounds, settled once the leader-runner-up gap exceeds 0.1; a Judge Agent breaks intra-group ties and a Consistency Agent reconciles cross-task conflicts. For joint extraction, the Consistency Agent runs one additional reassessment that drops or retypes flagged entities when its confidence falls below 0.75. All agents use temperature 0.9 and frequency penalty 0.1.

C Analysis and Discussion

C.1 Backbone Generalization

Table 10 evaluates whether the gains transfer to gemini-3-flash-preview. SMADE-IE achieves the best $F1_P$ and $F1_S$ on OntoNotes5, the best $F1_S$ on SemEval2010, and the best JERE results on CoNLL04, outperforming CrossAgentIE on CoNLL04 by 7.48 $F1_P$ and 4.84 $F1_S$. These results indicate that the framework generalizes across LLM backbones.

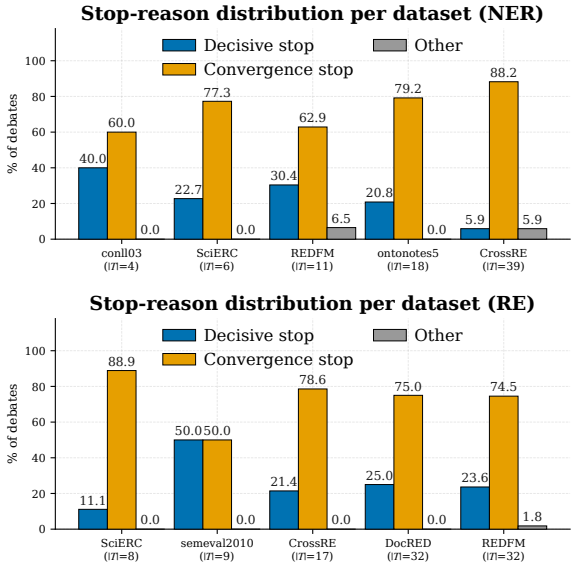

Figure 5: Early-stopping reason distributions of SMADE-IE on NER and RE datasets.

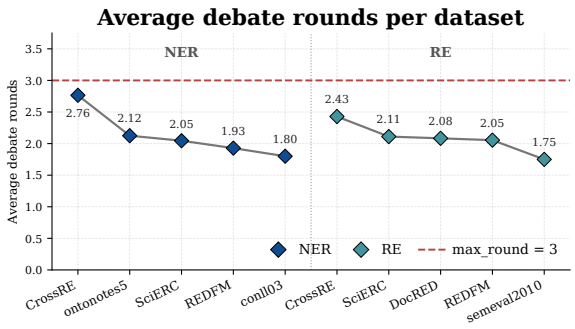

Figure 6: Average number of debate rounds under a maximum budget of 3.

C.2 Effectiveness of Early Stopping

Figures 5 and 6 report the early-stopping behavior of SMADE-IE. Confidence convergence is the dominant stopping reason, accounting for 60.0%–88.2% of early stops on NER and 74.5%–88.9% on most RE datasets, whereas decisive stopping is more prominent on CoNLL03 (40.0%) and SemEval2010 (50.0%). The *Other* category corresponds to cases in which the Ground and Warrant components of one or both debate agents are terminated. Consistent with these distributions, SMADE-IE requires only 1.75–2.76 rounds on average under a maximum budget of three. Simpler datasets such as CoNLL03 and SemEval2010 stop earliest at 1.80 and 1.75 rounds, while the more ambiguous CrossRE requires 2.76 rounds for NER and 2.43 for RE. These results show that early stopping suppresses redundant debate while preserving

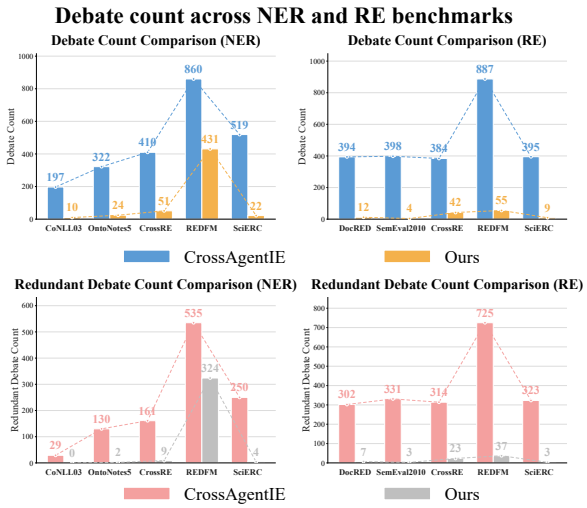

Figure 7: Debate and redundant-debate counts of CROSSAGENTIE and SMADE-IE on NER and RE datasets. Redundant debates refer to debates where the gold type does not exist among the proposed candidate types.

additional rounds for harder cases.

C.3 Debate Efficiency

Figure 7 reports a count-level analysis of the source of token savings. Across the ten NER and RE settings, SMADE-IE reduces the total number of debates from 4,766 to 660. For NER, substantial reductions are observed on every dataset: from 197 to 10 on CoNLL03, from 322 to 24 on OntoNotes5, from 410 to 51 on CrossRE, and from 519 to 22 on SciERC. REDFM remains the most challenging setting, yet the debate count is still nearly halved, from 860 to 431. A similar trend holds for RE, where the number of debates decreases from 394 to 12 on DocRED, from 398 to 4 on SemEval2010, from 384 to 42 on CrossRE, from 887 to 55 on REDFM, and from 395 to 9 on SciERC. Redundant debates are likewise reduced from 3,100 to 412 overall, indicating that Router-based type selection eliminates most unnecessary comparisons rather than merely shortening informative ones.

C.4 Hyperparameter Sensitivity

Figure 8 reports hyperparameter sensitivity on OntoNotes5. The default configuration achieves the best overall result with 66.08 partial-match F1 and 52.05 full-match F1. Varying δ_{stop} produces only moderate fluctuation: partial-match F1 stays within 63.85–66.08 and full-match F1 within 49.07–52.85, indicating that the posterior-superiority threshold is robust. The convergence threshold ε is more influential: the default value 0.02 outperforms both stricter and looser settings, with partial-match F1 dropping to 62.38 at 0.01 and 61.82 at 0.1. The parameter γ exhibits the clearest trade-off. A small value of 3 under-allocates debate and reduces F1 to 60.30/45.79, while an overly large value of 12 also hurts performance, reaching only 62.20/45.03. These results suggest that the default configuration balances evidence accumulation and early stopping, and that SMADE-IE remains robust under moderate perturbations.

D Posterior Superiority Bound for Dual-Track Termination

In this section, we derive the Chebyshev-type upper bound used in §2.4.3 and show how it motivates two complementary stopping criteria, posterior superiority and confidence convergence, which jointly govern the dual-track termination mechanism.

D.1 Setup and Notation

Consider two competing candidates with independent Beta posteriors $p_i \sim \text{Beta}(\alpha_i, \beta_i)$ for $i \in \{1, 2\}$. Their posterior mean and evidence mass are

$$\hat{p}_i \triangleq \frac{\alpha_i}{\alpha_i + \beta_i}, \qquad N_i \triangleq \alpha_i + \beta_i,$$

and the variance of p_i is

$$\sigma_i^2 \triangleq \text{Var}(p_i) = \frac{\hat{p}_i(1 - \hat{p}_i)}{N_i + 1}.$$

Assume candidate 1 currently has the larger posterior mean, i.e.,

$$\Delta \triangleq \hat{p}_1 - \hat{p}_2 > 0. \tag{31}$$

Our goal is to bound the probability of incorrect selection, $\mathbb{P}(p_2 > p_1)$, namely the probability that the current runner-up is in fact better than the leader. This quantity is the complement of the posterior-superiority probability $\mathcal{P}(p_1 > p_2)$ used in §2.4.3.

D.2 Chebyshev Bound on the Confidence Gap

Let $D \triangleq p_1 - p_2$. Independence of p_1 and p_2 gives

$$\mathbb{E}[D] = \Delta, \qquad \text{Var}(D) = \sigma_1^2 + \sigma_2^2.$$

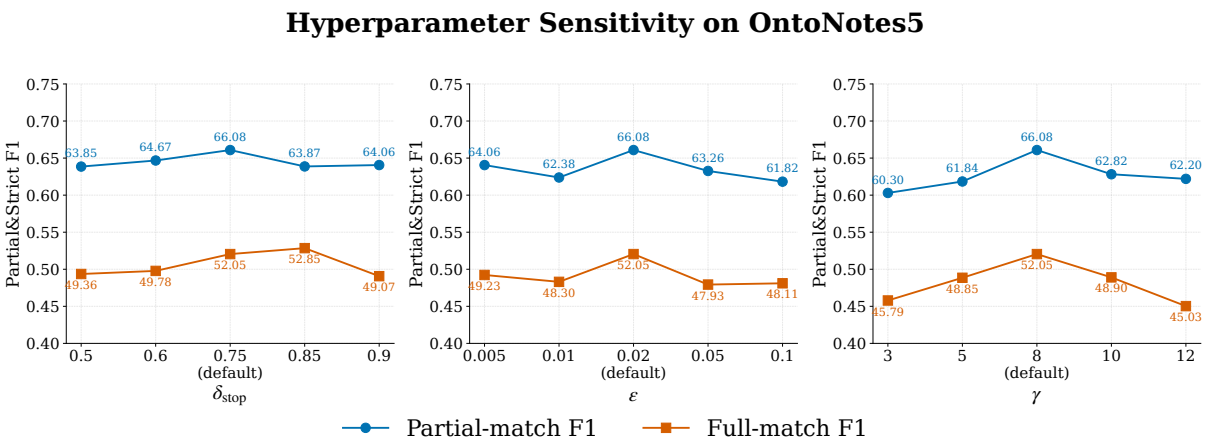

Figure 8: Hyperparameter sensitivity on OntoNotes5. We vary δ_{stop}, ε, and γ around their default values, where the default configuration ($\delta_{\text{stop}} = 0.75$, $\varepsilon = 0.02$, $\gamma = 8$) achieves the best overall F1.

The event $\{p_2 > p_1\}$ implies $\{|D - \Delta| \geq \Delta\}$. Applying Chebyshev's inequality yields

$$\begin{aligned} \mathbb{P}(p_2 > p_1) &\leq \frac{\mathrm{Var}(D)}{\Delta^2} \\ &= \frac{1}{\Delta^2}\left[\frac{\hat{p}_1(1-\hat{p}_1)}{N_1+1}\right. \\ &\quad \left. + \frac{\hat{p}_2(1-\hat{p}_2)}{N_2+1}\right]. \end{aligned} \tag{32}$$

The bound decreases as either the posterior gap increases or more evidence is accumulated.

D.3 Balanced-Evidence Special Case

When the two candidates have matched evidence mass, $N_1 = N_2 \triangleq N$, define the average posterior mean as $\mu \triangleq (\hat{p}_1 + \hat{p}_2)/2$, so that $\hat{p}_i = \mu \pm \Delta/2$. The numerator in Eq. (32) then becomes

$$\begin{aligned} f(\mu) &= \hat{p}_1(1-\hat{p}_1) + \hat{p}_2(1-\hat{p}_2) \\ &= -2\mu^2 + 2\mu - \tfrac{\Delta^2}{2}. \end{aligned} \tag{33}$$

This concave quadratic attains its maximum at $\mu = 1/2$, yielding the worst-case value $f(1/2) = \frac{1-\Delta^2}{2}$. Substituting this result into Eq. (32) gives the symmetric bound

$$\mathbb{P}(p_2 > p_1) \leq \frac{1}{2(N+1)}\left(\frac{1}{\Delta^2} - 1\right). \tag{34}$$

Under balanced evidence, the probability of incorrect selection is jointly controlled by two observable quantities: the posterior mean gap Δ and the evidence mass N.

D.4 Implications for Dual-Track Stopping

Eq. (34) reveals two complementary regimes under which the upper bound becomes small. Each regime corresponds to one stopping signal used in §2.4.3.

Posterior superiority. When the leading candidate is well separated from the alternatives, Δ becomes large and the bound in Eq. (34) decreases accordingly. The posterior-superiority criterion of Eq. (25) fires exactly in this regime: once $\mathcal{P}(\pi_{l^*}^{(t)} > \pi_{\bar{l}^*}^{(t)}) > \delta_{\text{stop}}$, the probability of incorrect selection is sufficiently suppressed, making additional debate unlikely to overturn the current decision.

Confidence convergence. When candidate types are semantically similar, Δ may remain small even after multiple rounds. In this regime, the bound can only be further reduced through additional informative evidence. The validity-decay rule in Eq. (18) progressively downweights repeatedly weakened components, causing the posterior updates accumulated by Eq. (19) to diminish over time. Consequently, once the posteriors become nearly unchanged across consecutive rounds, further debate contributes little additional information. This stabilization is exactly what the squared-Hellinger criterion of Eq. (22) measures: terminating once $\Delta_{\text{conf}}^{(t)} < \varepsilon$ matches the regime in which Eq. (34) cannot be tightened further without new evidence.

Together, posterior superiority addresses the regime of a dominant candidate, while confidence convergence captures the regime of diminishing posterior updates. The two signals jointly control the probability of incorrect selection without requiring a fixed debate budget.

E Case Study

We isolate the contribution of each SMADE-IE component through three contrastive cases decoded by the GPT-3.5-Turbo-0125 backbone. Each per-stage trace is contrasted with the final prediction of CROSSAGENTIE under the same backbone. The cases probe the Verification Agent on a context-dependent type mislabel in Listing 1, the Review Agent and Evidence-Driven Debate on a routed-out correct type in Listing 2, and the IERA loop on a joint extraction with spurious relations in Listing 3.

E.1 Case 1: Verification flips a context-dependent type mislabel

```
[SENTENCE]
In an interview with Blender (can't remember
what issue, but Janet 's on the cover), Janet
talks at length about having mulitiple
personalities.

[GOLD]
Blender -> ORG
Janet   -> PERSON

[SMADE-IE]
  -- NER stage --
  Router      : types = {PERSON, WORK_OF_ART},
                c = low  -> Global Extraction Mode
  Universal   : {Blender: WORK_OF_ART,
                 Janet:   PERSON}
  Verification: reject {Blender: WORK_OF_ART},
                add    {Blender: ORG}
  Final pred  = Gold                  Strict F1 = 1.00

[CrossAgentIE]
  All 18 OntoNotes types selected in parallel.
  Claims on "Blender": {PRODUCT, WORK_OF_ART}
                       -> 2-way conflict
                       (ORG agent silent).
  Debate    : free-form opinion exchange,
              summarizer tie-break -> WORK_OF_ART
              (the correct ORG label never enters
              the debate space).
  Spurious  : 'an interview' tagged as EVENT.
  Final pred = [Blender: WORK_OF_ART,
                Janet: PERSON,
                an interview: EVENT]
  Strict F1 = 0.571
```

Listing 1: OntoNotes5 NER sample 3621 trace.

The Universal Agent's literal reading misclassifies `Blender` as WORK_OF_ART, treating the magazine title as a creative work, and would have committed to the wrong label in a single call. Because the Verification Agent receives both the candidate set and the full sentence at once, it observes the discriminating context "interview with Blender" and resolves the conflict by simultaneously rejecting the WORK_OF_ART reading and adding the ORG reading in one call. In contrast, CROSSAGENTIE's all-agent claim phase admits only PRODUCT and WORK_OF_ART on this span, so the correct ORG label never enters the debate space; the GroupChat tie-break therefore settles on the wrong label, and the schema-wide selection additionally introduces a spurious EVENT span. Calibrated verification over a routed type subset, rather than schema-wide agent claims, enables a single additional call to recover a mistyped mention without inflating the candidate set.

E.2 Case 2: Review recovers a routed-out type and debate confirms it

```
[SENTENCE]
Panoramio was a geo-located tagging, photo
sharing mashup active between 2005 and 2016.
Photos uploaded to the site were accessible
as a layer in Google Earth and Google Maps.
The site's goal was to allow Google Earth
users to learn more about a given area by
viewing the photos that other users had taken
at that location. Panoramio was acquired by
Google in 2007. In 2009 the website was
among 1000 most popular websites worldwide.

[GOLD]
Panoramio      -> media
Google         -> organization
2005, 2016, 2007 -> date

[SMADE-IE]
  -- NER stage --
  Router      : types = {concept, date, organization},
                c = medium
                # 7 of 10 REDFM types pruned
                -> Type-Centric Extraction Mode
  Type-Specific Agents on routed types:
    organization agent -> Panoramio: organization
                          Google: organization
    date         agent -> 2005, 2016, 2007: date
    concept      agent -> []
  Review Agent (covers residual 7 types):
                      -> Panoramio: media
                      # 'media' was missed by Router
                      # but recovered here.
  Conflict detected on span "Panoramio":
                      {organization, media}
                      -> promoted to debate

  -- Debate stage on "Panoramio" --
  Init q   : media = 0.448,  org = 0.364   gap = 0.084
  Round 1 pi: media = 0.623,  org = 0.449   gap = 0.174
  Round 2 pi: media = 0.693,  org = 0.418   gap = 0.275
              stop via convergence track
  Winner = media                       matches Gold

[CrossAgentIE]
  Selects all 10 REDFM types in parallel.
  Claims on "Panoramio": {organization}
                         (media agent silent
                          on this span).
  No conflict raised -> commits to organization
  directly, off-gold.
  Spurious: Google Earth / Google Maps tagged as
            'event'/'location' from distractor
            agents.
  Final pred on "Panoramio" = organization
                                      (wrong)
```

Listing 2: REDFM NER sample 334 trace on the span `Panoramio`.

The Router selects only {concept, date, organization} for this sentence, leaving the correct type *media* in the residual set. The *organization* agent assigns "Panoramio" to its type based on the plausible company-like surface form, while the Review Agent, scanning the residual seven types, also recovers "Panoramio" under *media*. These two predictions conflict, and a binary debate be-

tween *media* and *organization* is launched with a narrow initial qualifier gap of 0.084. Two rounds widen the gap to 0.275 as the *media* side cites the sentence-internal mentions of "mashup", "photo sharing site", and "website" as Ground, attacks the organization-side Warrant on type fit, and converges. CROSSAGENTIE selects the full ten-type alphabet upfront, yet its *media* agent remains silent on this span, so no conflict is raised and the system commits to *organization* without scrutiny. This case shows that Review-driven candidate recovery and structured debate are complementary: Review reinstates the type the Router would otherwise exclude, and debate then promotes the recovered candidate over the routed-but-incorrect one using evidence that a uniform multi-agent vote would not surface.

E.3 Case 3: IERA reconciles a joint extraction with spurious relations

```
[SENTENCE]
The plane, owned by Bradley First Air, of
Ottawa, Canada, was carrying cargo to Montreal
for Emery Air Freight Corp., an air freight
courier service with a hub at the Dayton airport.

[GOLD]
(Bradley First Air, Org based in, Ottawa)
(Bradley First Air, Org based in, Canada)
(Ottawa,            Located in,    Canada)

[SMADE-IE]
  -- NER stage --
  Router      : {Loc, Org}, c = low
  Universal   : {Bradley First Air: Org,
                 Ottawa: Loc, Canada: Loc,
                 Montreal: Loc,
                 Emery Air Freight Corp.: Org,
                 Dayton airport: Loc}        (6)
  Verification: add {Dayton: Loc}             (+1)
  Entities    = 7 spans

  -- RE stage on the 7-entity set --
  Router      : {Org based in, Located in}, c = low
  Universal   :
    (Bradley First Air, Org based in, Ottawa)
    (Ottawa,            Located in,   Canada)
    (Emery Air Freight Corp., Org based in,
                              Dayton airport)
  Verification: add (Bradley First Air,
                     Org based in, Canada)
  Relations   = 4   # includes 1 spurious edge

  -- IERA pass 1 --
  RE rerun on the new entity set:
    Universal   : same 3 + (Bradley First Air,
                            Org based in, Canada)
    Verification: reject (Emery Air Freight Corp.,
                          Org based in,
                          Dayton airport)
  Relations   = 3   # spurious edge removed

  Final pred = Gold                  Strict F1 = 1.00

[CrossAgentIE]
  7 relations emitted:
    (Bradley First Air, Org based in, Ottawa)
    (Emery Air Freight Corp., Org based in, Montreal)
    (Emery Air Freight Corp., Org based in, Dayton)
    (Ottawa,            Located in,   Canada)
    (Montreal,          Located in,   Canada)
    (Dayton,            Located in,   Canada)
    (Dayton,            Located in,   Dayton airport)
  Strict F1 = 0.40
```

Listing 3: CoNLL04 Joint sample 48 trace.

Single-pass joint extraction over-emits relations on co-mentioned organizations and locations: the RE Universal Agent interprets "hub at the Dayton airport" as an ORGANIZATION BASED IN edge, while the NER step extracts every location-like span. IERA re-invokes RE under the ontology-reconciled entity set, at which point the Verification Agent rejects the spurious EMERY–DAYTON AIRPORT edge that no longer has supporting evidence, and the loop reaches a fixed point with three correct relations. CROSSAGENTIE lacks this consistency loop and emits five additional locative edges among the same co-mentioned spans. This loop is the mechanism that converts locally plausible candidates into globally consistent triples.

Summary. The three traces share a single mechanism: SMADE-IE allocates computation in proportion to instance difficulty. One additional call resolves a trivial omission, Review combined with debate rescues a routed-out type, and a short loop reconciles the joint extraction. CROSSAGENTIE applies a uniform all-type, all-debate budget and incurs noise from every distractor type it admits. Calibrated allocation, rather than larger ensembles, is what translates agent collaboration into reliable predictions.

F Prompts

In this section, we list the prompts used by every agent in SMADE-IE. System prompts are fixed; user prompts are instantiated by filling the angle-bracketed fields with instance-specific values. All agent outputs are returned as JSON and parsed with tolerance for minor formatting inconsistencies; on parse failure, the pipeline reverts to the safe default of the corresponding agent in §2. Dataset-specific type definitions are loaded from each dataset's ontology at runtime and are therefore elided from the templates below.

F.1 Router Agent Prompt

As formalized in Eq. (1), the Router Agent maps a sentence s and the entity type set $\mathcal{T}_e$ to a selected subset $\mathcal{A} \subseteq \mathcal{T}_e$ together with a complexity label $c \in \{\text{low}, \text{med}, \text{high}\}$. To preserve high recall at minimal overhead, the agent performs lightweight type selection without per-type rationales and consumes a single LLM call per sentence.

```
SYSTEM
You are a Router Agent for Entity Extraction.
Scan the text and identify which types are present.
Type Definitions:
{type_definitions}

Classify the input as low, medium, or high.
low:    explicit, unambiguous items with few clearly
      separated types.
medium: multiple types, mild ambiguity, fuzzy boundaries,
      or moderate context dependence.
high:   strong ambiguity, overlapping candidates, implicit
      context, or many interacting types.

For hierarchical overlaps, choose the most specific type.
List each type at most once regardless of how many
      instances of it appear in the text.

USER
Identify which types are present in the text.
Text: {sentence}

JSON: {"possible_types": ["<type>", ...],
       "complexity": "<low|medium|high>"}
```

For the RE instantiation, "Entity Extraction" is replaced by "Relation Extraction" and the type definitions are drawn from the relation ontology.

F.2 Universal Agent Prompt

Eq. (2) specifies the Universal Agent, which serves the Global Extraction Mode on low-complexity inputs and extracts, in a single LLM call, all candidates whose types lie in the selected subset $\mathcal{A}$.

```
SYSTEM
You are a Universal Entity Extraction Agent covering all
      types.
Extract the exact noun phrase as the span. Do not include
      determiners (the, a, an) or unnecessary adjectives.
An entity can only belong to one category without overlap.

Type Definitions:
{type_definitions}

USER
Extract all entities whose type is one of:
      {selected_types}. Return empty dict if none.

Text: {sentence}

JSON: {"<verbatim span>": "<type>", ...}
```

For RE, the JSON template becomes `{"<head>|||<tail>": "<type>", ...}`.

F.3 Review Agent Prompt

Eq. (6) defines the Review Agent, which operates under the Type-Centric Extraction Mode on medium- and high-complexity inputs. It covers the residual type set $\mathcal{A}^* = \mathcal{T}_e \setminus \mathcal{A}$, recovering candidates whose types lie outside the Router-selected subset and compensating for any Router omission. The Review Agent reuses the system prompt of the Universal Agent and adopts the user prompt below, in which `remaining_types` instantiates $\mathcal{A}^*$.

```
USER
Extract entities whose type is only in {remaining_types}.
      Assign exactly one type per entity; return [] if none.
Text: {sentence}

JSON: {"results": [{"span": "<verbatim text>",
                    "type": "<type>"}, ...]}
```

For RE, the JSON template replaces `"span"` with the two fields `"head"` and `"tail"`.

F.4 Verification Agent Prompt

Per Eq. (3), the Verification Agent is invoked only on the low-complexity path, after Universal Agent extraction. Given the candidate set, it removes incorrectly typed candidates and inserts missed entities in a single LLM call, improving extraction quality while keeping the entire path within two calls.

```
SYSTEM
You are a Verification Agent for Entity type verifications.
Extract the exact noun phrase as the span. Do not include
      determiners (the, a, an) or unnecessary adjectives.
If a candidate span contains multiple independent entities
      separated by punctuation, reject the combined span
      and add each individual entity separately.
Type Definitions:
{type_definitions}

USER
Text: "{sentence}"

Candidates:
"{span_1}" is {type_1}
"{span_2}" is {type_2}
...

Reject candidates with incorrect types and add any missing
      entities from the text. The type must be one of the
      defined types.

JSON: {"added":    [{"span": "<verbatim text>", "type":
      "<type>"}],
       "rejected": [{"span": "<verbatim text>", "type":
      "<type>"}]}
```

For RE, each candidate is rendered as a triple `(<head>, <type>, <tail>)`, and the JSON fields replace `"span"` with `"head"` and `"tail"`.

F.5 Type-Specific Agent and Toulmin-Guided Argument Construction Prompt

Each Type-Specific Agent is bound to a single candidate type $t^i_{\mathcal{A}} \in \mathcal{A}$ and operates in two phases. The extraction phase, governed by Eq. (5), scans the sentence for spans of the bound type and contributes them to the candidate set $E_{\mathcal{A}}$ of the Type-Centric Extraction Mode. When a candidate entity $\tilde{e}^i$ is later contested by multiple types $\{\tilde{t}^i_1, \ldots, \tilde{t}^i_N\}$, the agent for type $\tilde{t}^i_n$ enters the debate phase and emits the Toulmin argument $\mathcal{D}_n = \{C_n, G_n, W_n, B_n, R_n, Q_n\}$ specified by Eq. (9). The qualifier Q_n is not produced by the agent itself but is computed externally by the scorer ϕ over the argument paragraph; the rebuttal R_l of each surviving candidate l is reused pre-debate to set its prior strength κ_l via the rebuttal-honesty derivation of

Eq. (12).

The extraction phase uses the per-type system and user prompts:

```
SYSTEM
You are an Entity extraction Agent for type: {agent_type}.
Definition of {agent_type}: {type_definition}
Extract the exact noun phrase as the span. Do not include
      determiners (the, a, an) or unnecessary adjectives.
An entity can only belong to one category without overlap.

USER
Extract all "{agent_type}" entities from the text. Return
      empty list if none.

Text: {sentence}

JSON: ["<span1>", "<span2>", ...]
```

In the debate phase, the same agent is reinstantiated under a debate-specific system message that supplies the contested span, the sentence, and all competing type definitions. The Toulmin-generation user prompt is then issued once per conflict:

```
SYSTEM
You are {agent_type} Agent in a debate over the entity
      "{span}".
Sentence: {sentence}

Type definitions:
- {type_1}: {def_1}
- {type_2}: {def_2}
...

In each round you receive an opponent's textual evidence
      Ground and type-fit reasoning Warrant. Write counters
      that both support {agent_type} and rebut the
      opponent's claim. If your own components get attacked
      enough, you will be asked to revise them.

USER
Conflict detected on span: "{span}" in "{sentence}"

Construct a Toulmin argument as JSON:
{
  "Claim":    "The correct type for {span} is {agent_type}",
  "Ground":   "<factual evidence quoted from the text>",
  "Warrant":  "<reasoning rule linking Ground to Claim>",
  "Backing":  "<additional linguistic or domain support>",
  "Rebuttal": "<one concise sentence that presents a
      counter-view conflicting with the Claim, quoting an
      original sentence fragment as evidence; no
      conditionals, no hedging>"
}
```

F.6 Refutation Prompt

In each debate round, the opposing agent $\mathcal{D}_{\bar{l}}$ attacks the defended agent $\mathcal{D}_l$ over the active subset $\mathcal{V}_l^{(t)} \subseteq \{G, W\}$ defined by Eq. (14). A single LLM call yields one refutation $u_{\bar{l},m}$ per defended component $m \in \mathcal{V}_l^{(t)}$; each refutation is scored against the defender by the evidence scorer ϕ and then folded into the Beta-posterior update of Eq. (19).

```
USER
Counter the {target_type} Agent's argument.
Ground:  cite a fragment from the text that supports
      {attacker_type} and shows the opponent's evidence
      does not establish {target_type}.
Warrant: argue from the type definitions that
      {attacker_type} fits the entity and {target_type}
      does not.
Take a new angle if you have replied earlier.

Opponent's Ground:  {ground_text}
Opponent's Warrant: {warrant_text}

JSON: {"Ground":  "<counter-claim>",
       "Warrant": "<counter-claim>"}
```

When only Ground or only Warrant remains in $\mathcal{V}_l^{(t)}$, the instruction line, opponent excerpt, and JSON field corresponding to the inactive component are omitted; the per-round LLM budget therefore scales with the number of attackable components rather than with their nominal count.

F.7 Revision Prompt

When the decay rule of Eq. (18) drives one or more components of an agent's argument to a validity at or below the activity threshold θ_v, the owning agent must revise them in the next round. All revisions belonging to the same agent are batched into a single LLM call, so the per-round revision cost scales with the number of agents rather than with the total number of components to revise.

```
USER
Current Toulmin argument:
Claim:    "..."
Ground:   "..."
Warrant:  "..."
Backing:  "..."
Rebuttal: "..."

Your {forfeited_components} were attacked. Revise with
      stronger arguments.

JSON: {"<comp_1>": "<revised text>",
       "<comp_2>": "<revised text>"}
```

The JSON output contains exactly the components listed in `forfeited_components`; non-forfeited components are not regenerated.

F.8 Iterative Entity–Relation Alignment Prompt

As described in §2.5, the Consistency Agent is invoked whenever a triple in R contains a head or tail whose entity type in E violates $\mathrm{HeadTypes}(r)$ or $\mathrm{TailTypes}(r)$. All conflicts detected within the same sentence are batched into a single LLM call. The agent returns one of two decisions per conflict: under `NER_correct`, the offending relation is removed from R and added to the blacklist B, preventing any subsequent RE recomputation from re-introducing it; under `RE_correct`, the offending entity type in E is replaced by the constraint-compatible type returned by the agent.

```
SYSTEM
You are a consistency arbitrator for NER--RE conflicts.
When an entity's type contradicts a relation's type
      constraint, decide which side is correct.
```

```
Entity Ontology:
{entity_definitions}

Relation Ontology:
{relation_definitions}

USER
Text: "{sentence}"

You have {N} conflict(s) to resolve:

Conflict 1:
  Entity:   "{entity_span_1}" typed as "{entity_type_1}" by
      NER
  Relation: ({head_span_1}, {relation_type_1},
      {tail_span_1})
  Role:     {role_1}
  Expected types: [{expected_types_1}]
...

For each conflict, decide:
- "NER_correct": the entity type is correct; remove the
      relation.
- "RE_correct":  the relation is correct; replace the
      entity type with one of the expected types.

JSON: {"decisions": [{"conflict_id":    <1-based index>,
                      "decision":
      "<NER_correct|RE_correct>",
                      "corrected_type": "<type if
      RE_correct, else empty>"}, ...]}
```

F.9 Span Classification for RE → NER Back-Fill Prompt

The span-classification step of §2.5 assigns an entity type to every span that appears as a head or tail in R but is absent from E. The Universal Agent is repurposed in span-classification mode with the user prompt below; its system prompt is identical to the one used by the Universal Agent elsewhere in this appendix.

```
USER
Please classify the entity type for each span below. Use
      only the defined types.

Text: {sentence}
Spans to classify: ["{span_1}", "{span_2}", ...]

JSON: {"classifications": [{"span": "<verbatim text>",
                            "type": "<type>"}, ...]}
```