

Reflect then Learn: Active Prompting for Information Extraction Guided by Introspective Confusion

Dong Zhao^{1*}, Yadong Wang^{1*}, Xiang Chen^{1†}, Chenxi Wang², Hongliang Dai¹, Chuanxing Geng¹, Shengzhong Zhang¹, Shao-Yuan Li¹, Sheng-Jun Huang^{1†},

¹MIT Key Laboratory of Pattern Analysis and Machine Intelligence,
College of Computer Science and Technology,
Nanjing University of Aeronautics and Astronautics
²Zhejiang University
{zhao.dong, xiang_chen}@nuaa.edu.cn

Abstract

Large Language Models (LLMs) show remarkable potential for few-shot information extraction (IE), yet their performance is highly sensitive to the choice of in-context examples. Conventional selection strategies often fail to provide informative guidance, as they overlook a key source of model fallibility: confusion stemming not just from semantic content, but also from the generation of well-structured formats required by IE tasks. To address this, we introduce **Active Prompting for Information Extraction (APIE)**, a novel active prompting framework guided by a principle we term *introspective confusion*. Our method empowers an LLM to assess its own confusion through a dual-component uncertainty metric that uniquely quantifies both Format Uncertainty (difficulty in generating correct syntax) and Content Uncertainty (inconsistency in extracted semantics). By ranking unlabeled data with this comprehensive score, our framework actively selects the most challenging and informative samples to serve as few-shot exemplars. Extensive experiments on four benchmarks show that our approach consistently outperforms strong baselines, yielding significant improvements in both extraction accuracy and robustness. Our work highlights the critical importance of a fine-grained, dual-level view of model uncertainty when it comes to building effective and reliable structured generation systems.

Code — <https://github.com/NUAA-MMMI/APIE>

Introduction

Information Extraction (IE) serves as a foundational pillar for converting unstructured text into structured, machine-readable knowledge, thereby empowering a multitude of downstream applications such as knowledge graph construction, semantic search, and intelligent question answering (Pai et al. 2024; Ma et al. 2023; Chen et al. 2022a). Traditionally, state-of-the-art IE systems have been dominated by fully supervised paradigms which necessitate extensive

*These authors contributed equally.

†Corresponding authors.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

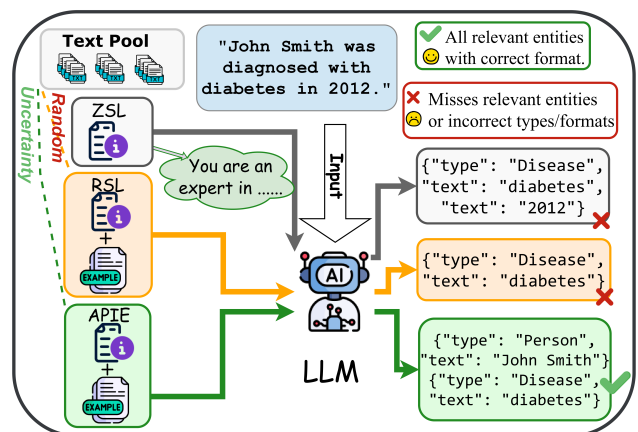


Figure 1: Illustration of Prompt Construction and Entity Extraction for ZSL, RSL, and APIE. Our approach achieves consistently higher extraction accuracy by selecting high-quality examples via structured uncertainty scoring, while ZSL and RSL use fixed or random examples, leading to less accurate results across all domains.

domain-specific labeled data for training. This reliance on large-scale annotation renders them costly and difficult to scale, particularly in specialized domains like medicine, law, or finance where expert annotation is a significant bottleneck and beyond. The emergence of Large Language Models (LLMs) has catalyzed a paradigm shift toward few-shot learning, where models perform complex tasks by conditioning on a small number of in-context examples (Brown et al. 2020). This prompt-based approach dramatically reduces the dependency on labeled data and offers unprecedented flexibility. However, the efficacy of LLMs in structured IE tasks is contingent on the quality of the provided exemplars. Prevailing selection strategies such as random sampling or simple semantic similarity often prove sub-optimal and lead to brittle performance as they fail to furnish the model with the most informative guidance for the intricate demands of IE and robustness (Margatina et al. 2023).

This sensitivity stems from a fundamental - yet often overlooked - dichotomy in the challenges inherent to structured generation tasks with LLMs. The model’s fallibility arises not merely from semantic ambiguity in the source text, but also from the complex task of adhering to rigid output schemas. Existing uncertainty-guided prompting methods, while effective for classification tasks with discrete label spaces (Diao et al. 2024), fall short in this context. They typically overlook a crucial dimension of model confusion: the struggle to maintain syntactic fidelity (i.e., structural correctness and parsability of the output) while simultaneously resolving semantic uncertainty (i.e., the correctness of the extracted content). This dual challenge of format and content necessitates a more nuanced approach to example selection.

To bridge this critical gap, we introduce APIE, an uncertainty-driven, training-free prompting framework designed for universal information extraction. Our approach is guided by a principle we term introspective confusion, which empowers the LLM to “reflect” on its own generative process and identify samples that are most challenging and thus most valuable for learning. The model’s internal confusion can be measured by analyzing the syntactic and semantic consistency across its own multiple, independently generated outputs. Specifically, we propose a dual-level introspective uncertainty metric that quantifies this confusion from two critical angles: 1) **Format-Level Uncertainty** which captures the model’s struggle to produce structurally coherent and parsable outputs, measured through a combination of parsing failures and generation disagreement; and 2) **Content-Level Uncertainty** which assesses the semantic consistency of extracted information across multiple inferences using set-based divergence.

By ranking unlabeled data with this comprehensive uncertainty score, APIE actively selects exemplars that are both structurally complex and semantically ambiguous. As illustrated in Figure 1, unlike zero-shot (ZSL) or random-sample (RSL) prompting which often lead to format errors or incorrect extractions, APIE constructs highly informative prompts that guide the LLM toward producing accurate and well-structured outputs. We validate our framework through extensive experiments on four diverse IE benchmarks, demonstrating consistent and significant improvements over strong baselines across multiple LLMs. Our main contributions are summarized as follows:

- A novel active prompting framework, **APIE**, that uniquely introduces the principle of introspective confusion to better address the dual challenges of format and content generation in structured IE.
- The formulation of a dual-level introspective uncertainty estimation mechanism that uniquely combines format-level instability (parsing failures, generation variance) and content-level divergence (semantic inconsistency) to guide the selection of high-utility exemplars.
- Extensive empirical validation across four benchmark datasets and four LLMs, demonstrating that APIE confidently establishes a new state-of-the-art for training-free universal information extraction, with significant gains in both overall accuracy and robustness.

Related Work

Prompt-based Information Extraction with LLMs

The advent of large language models has profoundly reshaped the landscape of information extraction (IE), shifting the focus from task-specific fine-tuning to versatile, prompt-based frameworks (Chen et al. 2022b; Liu et al. 2023). These methods reframe IE as a language modeling problem, where a carefully crafted prompt guides an LLM to generate structured outputs for tasks like named entity recognition (NER), relation extraction (RE), and event extraction (EE) (Xie et al. 2022; Lu et al. 2022). This paradigm includes both zero-shot and few-shot prompting.

Despite their promise, the performance of these methods is highly sensitive to prompt design, especially the selection of in-context examples (Diao et al. 2024; Han et al. 2023). Common strategies, such as random sampling or k-NN based on embedding similarity (Wang et al. 2025), often fail to capture the nuanced structural and semantic complexities of IE tasks, resulting in unstable performance and a high variance across different example sets (Liu et al. 2022). To mitigate issues with output structure, some works have explored schema-guided prompting or post-processing parsers to enforce format constraints. However, these methods reactively correct or constrain the output, rather than proactively improving the model’s intrinsic understanding of the task. They lack an adaptive selection mechanism that identifies exemplars capable of resolving the model’s own confusion, a critical gap our work aims to fill.

Uncertainty-Guided Prompt Optimization

Uncertainty estimation is a cornerstone of active learning, where it is used to identify the most informative samples for labeling, thereby maximizing model performance with minimal annotation effort (Lewis and Gale 1994; Gal, Islam, and Ghahramani 2017; Settles 2009). Recently, these principles have been adapted to the LLM prompting paradigm. For instance, Active-Prompt (Diao et al. 2024) leverages uncertainty metrics like generation disagreement and entropy to select effective exemplars for complex reasoning tasks. Similarly, other studies have explored uncertainty to automate prompt engineering or select diverse demonstrations (Margatina et al. 2023; Qian et al. 2024; Zhang et al. 2025).

However, these pioneering efforts target classification or multi-choice reasoning tasks, where the output space is discrete and predefined. Their uncertainty metrics are ill-suited for the open-ended and highly structured nature of IE outputs (Diao et al. 2024). The challenge in IE is not just choosing the correct label from a list, but instead involves generating a complex, composite structure with both correct content and valid syntax. Existing methods do not account for the model’s uncertainty in generating a well-formed structure, a frequent point of failure for LLMs in IE.

In contrast, APIE introduces a dual-level uncertainty framework tailored specifically for structured generation. While prior works measure disagreement at the final-answer level, our approach decomposes uncertainty into two orthogonal dimensions: syntactic fidelity and semantic uncertainty. This allows us to distinguish between a model that is con-

fused about the content and one that is confused about the format. APIE is the first framework to leverage such a holistic view of model confusion to actively and adaptively construct prompts for universal information extraction.

Methodology

Our proposed framework, APIE, introduces a novel active prompting strategy that selects high-utility exemplars by quantifying and leveraging the model’s own introspective confusion. This section begins by formulating the task of universal information extraction as a prompt optimization problem under the few-shot learning setting. We then provide a comprehensive overview of the APIE pipeline, followed by a detailed exposition of its core components: a dual-level uncertainty estimation mechanism and a structured, uncertainty-guided prompt construction process tailored for maximizing information gain.

Problem Formulation

We formulate information extraction (IE) as the task of learning a mapping function $\mathcal{F} : x \rightarrow y$, where x is a piece of unstructured text from a domain \mathcal{X} , and y is its corresponding structured representation (e.g., a set of entities, relations, or events) from a schema \mathcal{Y} . In a training-free, few-shot setting with an LLM \mathcal{M} , this mapping is conditioned on a prompt P that contains task instructions and a set of n exemplars $S = \{(x_j, y_j)\}_{j=1}^n$. The goal is to find an optimal set of exemplars S^* from an unlabeled data pool \mathcal{D}_u that maximizes the extraction performance on a test set:

$$S^* = \arg \max_{S \subseteq \mathcal{D}_u, |S|=n} \mathbb{E}_{(x_i, y_i) \sim \mathcal{D}_{test}} [\mathcal{E}(\mathcal{M}(x_i, P(S)), y_i)], \quad (1)$$

where $P(S)$ denotes the prompt constructed with exemplar set S , and \mathcal{E} is a task-specific evaluation metric (e.g., F1-score). Directly solving this combinatorial optimization problem is intractable. Thus, APIE proposes a principled proxy: selecting exemplars that maximize the model’s estimated uncertainty as a surrogate objective.

Output Schema and Task Setup

To ensure consistency and enable automatic evaluation, we define a unified structured output schema strictly enforced for all IE tasks. The model’s output must be a parsable JSON list of objects, where each object represents an extracted piece of information. For entities, objects must contain "type" and "text" keys; for relations, they require "type", "head", and "tail". Our strict parser, \mathcal{R} , validates each generation against this schema; any deviation, such as malformed JSONs or missing required keys, is flagged as a format failure and contributes directly to the Format-Level Uncertainty (\mathcal{U}_f) calculation.

We also establish rigorous task boundaries for evaluation. Models are expected to produce an empty list ($[]$) for sentences containing no extractable information. For evaluation, we employ a strict matching protocol, where both the text span and type of an extracted element must exactly match the ground truth; partial matches are considered incorrect. This protocol is applied consistently across all methods

for the final F1-score computation to ensure a fair and direct comparison without any exception whatsoever.

The Overview of our Framework

The APIE framework operationalizes the principle of Introspective Confusion through a systematic three-stage pipeline, illustrated in Figure 2. First, the framework probes an unlabeled pool by prompting the LLM to produce diverse outputs for each sample. These generations are then processed by our uncertainty metric, which integrates surface-level disagreement \mathcal{U}_d with format-level \mathcal{U}_f and content-level \mathcal{U}_c scores into a unified measure of confusion. The pool is ranked accordingly, and the top high-uncertainty samples are selected as exemplars and placed into a pattern-constrained prompt with task instructions. This actively constructed prompt is then used to guide the LLM on new instances, yielding accurate and robust extractions.

Generation Disagreement

As a foundational baseline signal, we first measure the surface-level disagreement across k generated outputs $\{\mathcal{M}_\theta^j(s_i)\}_{j=1}^k$ for a given sample s_i . We then quantify this using the average pairwise Levenshtein distance, which captures variations in the raw text output:

$$\mathcal{U}_d(s_i) = \frac{2}{k(k-1)} \sum_{1 \leq j < l \leq k} \text{Levenshtein}(\mathcal{M}_\theta^j(s_i), \mathcal{M}_\theta^l(s_i)), \quad (2)$$

where a high \mathcal{U}_d value signals a general state of overall output inconsistency without distinguishing between format and content discrepancies. In this sense, \mathcal{U}_d can be viewed as a form of pre-format content uncertainty, capturing semantic variation prior to structural normalization.

Dual-Level Introspective Uncertainty

This is the core of our estimation strategy. Recognizing that failures in structured IE arise from distinct sources, we disentangle the model’s confusion into two orthogonal dimensions. This decomposition allows us to move beyond monolithic uncertainty scores and gain a more granular understanding of why a model struggles with a particular sample, enabling a more targeted selection of exemplars.

Format-Level Uncertainty. This metric quantifies the model’s difficulty in adhering to the required output syntax. We define it as a combination of two signals:

- **Parsing Failure Rate (R_{fail}):** The fraction of k outputs that fail to be parsed by a strict structure parser \mathcal{R} . A high failure rate indicates the model is fundamentally struggling to generate syntactically valid structures. The parsing failure rate is calculated as follows:

$$R_{fail}(s_i) = \frac{1}{k} \sum_{j=1}^k \mathbb{I}[\mathcal{R}(\mathcal{M}_\theta^j(s_i)) = \text{FAIL}], \quad (3)$$

where \mathbb{I} is the indicator function.

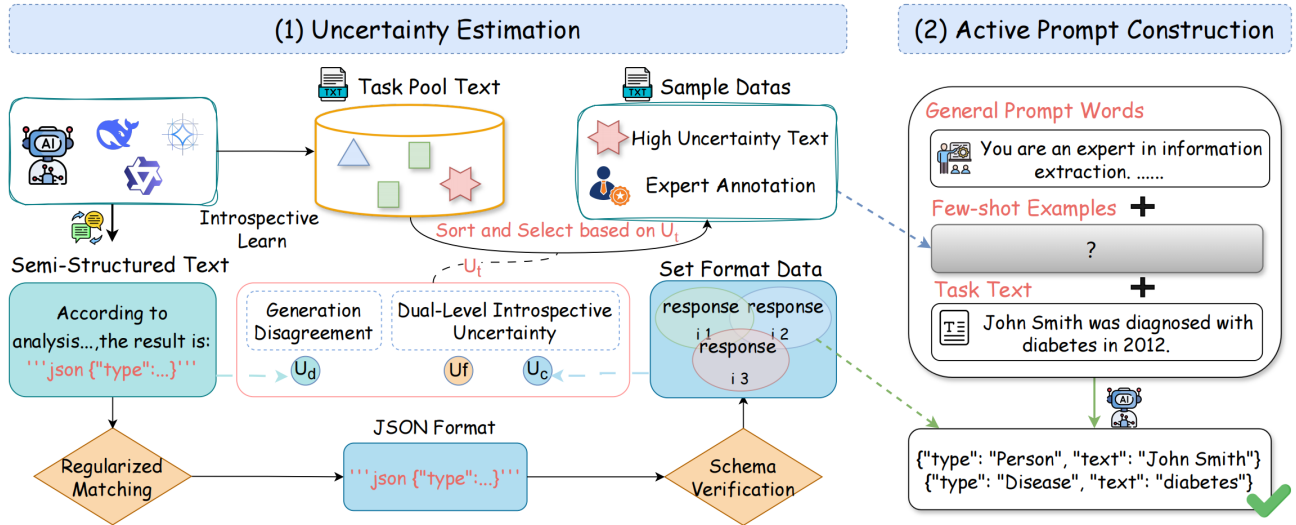


Figure 2: APIE consists of two main stages. (1) Uncertainty Estimation: Given an unlabeled pool, candidate exemplars are evaluated using three complementary uncertainty signals—disagreement-level (\mathcal{U}_d), format-level (\mathcal{U}_f), and content-level (\mathcal{U}_c)—to compute a unified uncertainty score \mathcal{U}_t and select high-utility examples. (2) Active Prompt Construction: The selected examples are incorporated into a structured prompt with task instructions and format guidance, enabling the language model to produce accurate, schema-conforming structured outputs (e.g., entity tuples). This framework enhances the informativeness and reliability of few-shot prompting without requiring labeled data.

- **Structural Disagreement:** For the successfully parsed outputs, we measure the variance in their structural composition (e.g., different sets of keys, varying list lengths). This captures more subtle inconsistencies in formatting.

The format-level uncertainty \mathcal{U}_f , a weighted sum of these components, is designed to directly measure the model’s struggle with syntactic fidelity.

Content-Level Uncertainty. This metric assesses the semantic consistency of the information extracted in the successfully parsed outputs. Even if an LLM generates perfectly formatted outputs, it may still be uncertain about the actual content such as the boundaries of extracted spans or the correctness of predicted types and relations. For each sample s_i , we obtain a set of k' valid structured outputs $B_i = \{b_1, b_2, \dots, b_{k'}\}$, where each b_j is a set of extracted (type, text) tuples. We define content-level uncertainty as one minus the average pairwise Jaccard similarity, which measures the overlap of extracted information:

$$\mathcal{U}_c(s_i) = 1 - \frac{2}{k'(k'-1)} \sum_{1 \leq j < l \leq k'} \frac{|b_j \cap b_l|}{|b_j \cup b_l|}. \quad (4)$$

A low Jaccard similarity (high \mathcal{U}_c) signifies that the model is generating semantically divergent extractions, indicating high uncertainty about the content. This signal is vital for identifying samples with inherent semantic ambiguity.

Active Prompt Construction

The final stage of our framework synthesizes the distinct uncertainty signals to construct a high-efficacy prompt. To ensure each signal contributes both fairly and appropriately to

the final score, we first normalize the disagreement (\mathcal{U}_d), format (\mathcal{U}_f), and content (\mathcal{U}_c) uncertainties to a common $[0, 1]$ scale. We then compute a unified uncertainty score, \mathcal{U}_{total} , for each candidate sample s_i by taking a weighted combination of these normalized facets:

$$\mathcal{U}_{total}(s_i) = \alpha \mathcal{U}_d(s_i) + \beta \mathcal{U}_f(s_i) + \gamma \mathcal{U}_c(s_i), \quad (5)$$

where α , β , and γ are **key** hyperparameters that weigh the contribution of each uncertainty component. This design provides a high degree of flexibility, allowing for the prioritization of different uncertainty types depending on the specific task demands.

Samples from the pool \mathcal{D}_u are ranked by their \mathcal{U}_{total} score and the top- n are selected. These selected texts, representing the most informative and challenging cases for the model, are then provided to a human expert for annotation to create the ground-truth labels. This active selection process significantly reduces the annotation burden as experts only need to label a small, targeted subset of data rather than a large random one, saving considerable human cost. The resulting high-quality exemplars are then seamlessly integrated into our prompt template to facilitate a more effective form of in-context learning, directly guiding the model toward more accurate and robust structured predictions.

Experiments

Experimental Setup

Tasks and Datasets. Our evaluation is conducted on four widely-used benchmarks spanning two core IE tasks: named entity recognition (NER) and relation extraction (RE). This

Dataset	$ Ent $	$ Rel $	#Test
ACE04	7	–	812
CoNLL03	4	–	3,453
CoNLL04	4	5	288
SciERC	6	7	551

Table 1: Detailed dataset statistics. Here, $|*|$ indicates the exact number of categories, and # is the exact number of sentences in the specific subset.

diverse set, summarized in Table 1, includes both single-task (CoNLL03 (Tjong Kim Sang and De Meulder 2003), ACE04 (Mitchell et al. 2005)) and joint NER-RE settings (CoNLL04 (Roth and Yih 2004), SciERC (Luan et al. 2018)), which allows for a thorough assessment of framework generalizability. We adopt an end-to-end setting where models receive only raw text and must generate the target structures from it directly without any preprocessing.

Models. To evaluate performance across varying capabilities and architectures, we select four publicly available, open-source LLMs: Gemma-3-12B (Team et al. 2025), Qwen-2.5-14B (Qwen et al. 2025), DeepSeek-R1-14B (DeepSeek-AI et al. 2025a), and DeepSeek-V3-660B (DeepSeek-AI et al. 2025b). This selection covers models with diverse architectures and scales, enabling a robust analysis of our method’s scalability.

Baselines. We compare APIE against four representative prompting strategies:

- **Zero-Shot (ZSL):** The model performs extraction based only on the given task instructions, without the guidance of any in-context examples.
- **Random-Sample (RSL):** A standard few-shot baseline where exemplars are randomly selected.
- **KD Sort:** A curriculum-inspired strategy that selects exemplars based on knowledge density, defined as entities/relations per unit text length.
- **Active-Prompt:** An uncertainty-guided baseline that uses output disagreement on final answers to select exemplars, but lacks the structured uncertainty analysis.

Evaluation Metrics. Following standard practice, we use micro-averaged F1-score as the primary metric for all tasks. To ensure reproducibility, all reported results are averaged over multiple independent runs with different random seeds.

Main Results

For RQ1: Does APIE consistently outperform established baselines in terms of extraction accuracy and robustness across different tasks and models? Our experiments demonstrate that APIE consistently outperforms all baselines in both accuracy and robustness. Regarding accuracy, the main results in Table 2 show that our framework achieves the highest average F1-score across nearly all backbones. The performance gains are particularly pronounced on complex joint-extraction tasks. For in-

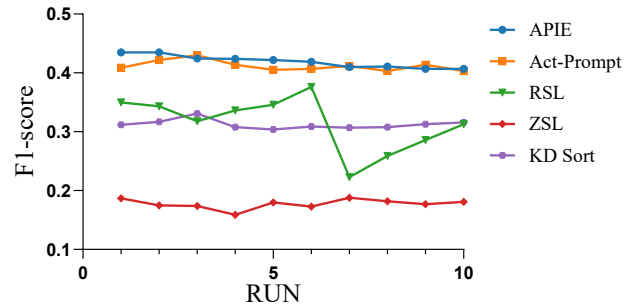


Figure 3: Performance stability comparison on the ACE04-NER dataset across 10 independent runs using the DeepSeek-R1-14B backbone.

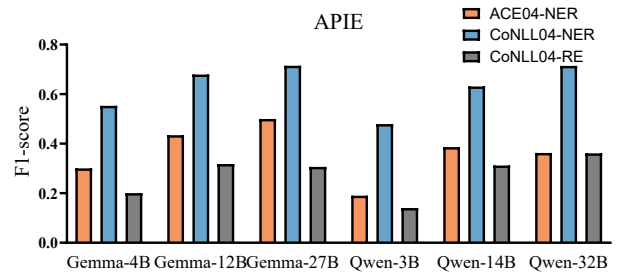


Figure 4: F1-score comparison across different model scales using APIE on three different datasets. Here, Gemma refers to Gemma-3, and Qwen refers to Qwen-2.5.

stance, with the DeepSeek-R1-14B model on the SciERC dataset, APIE achieves a 13.58 F1-score in relation extraction, marking a 18.3% relative improvement over the Active-Prompt baseline (11.48). This highlights the superior ability of our structured uncertainty metrics to identify informative exemplars for complex schema. Beyond superior accuracy, APIE also delivers greater robustness. As visualized in Figure 3, the performance of the Random-Sample (RSL) baseline fluctuates dramatically across runs, underscoring its unreliability. In contrast, APIE produces consistent results with minimal variance, confirming that our approach constructs stable and generalizable prompts.

For RQ2: How does the effectiveness of APIE scale with the capacity of the backbone large language models? We investigate the impact of model scale on performance with results shown in Figure 4. Our analysis reveals that APIE’s advantage is most pronounced on smaller-scale models such as Gemma-3-4B and Qwen-2.5-3B, where performance gaps against strong baselines exceed 10 F1 points on tasks such as ACE04-NER. This suggests that our structured uncertainty-guided prompting compensates effectively for the limited capacity and reasoning ability of compact LLMs. As model size increases, the overall performance of all prompting strategies improves, and inter-method performance gaps gradually narrow, indicating that large-capacity models are inherently more resilient to sub-

Models	Methods	CoNLL03	ACE04-NER	CoNLL04		SciERC		Average F1-score↑
		NER F1-score↑	NER F1-score↑	NER F1-score↑	RE F1-score↑	NER F1-score↑	RE F1-score↑	
Gemini-3-12B	ZSL	58.33 ↑0.00	22.48 ↑0.00	68.47 ↑0.00	20.59 ↑0.00	24.12 ↑0.00	14.53 ↑0.00	34.76 ↑0.00
	RSL	62.96 ↑4.62	39.76 ↑17.28	68.29 ↓0.18	26.92 ↑6.32	29.57 ↑5.45	14.43 ↓0.11	40.32 ↑5.56
	KD Sort	59.80 ↑1.47	40.11 ↑17.63	68.58 ↑0.11	36.71 ↑16.11	27.25 ↑3.13	9.84 ↓4.69	40.38 ↑5.63
	Active-Prompt	58.91 ↑0.57	43.61 ↑21.13	65.76 ↓2.71	28.38 ↑7.78	25.45 ↑1.33	19.25 ↑4.72	40.23 ↑5.47
	APIE	60.07 ↑1.73	44.42 ↑21.93	68.67 ↑0.20	28.82 ↑8.23	26.21 ↑2.09	16.85 ↑2.32	40.84 ↑6.09
Qwen-2.5-14B	ZSL	49.15 ↑0.00	10.78 ↑0.00	57.61 ↑0.00	13.69 ↑0.00	20.51 ↑0.00	18.08 ↑0.00	28.31 ↑0.00
	RSL	63.35 ↑14.20	30.05 ↑19.27	63.35 ↑5.74	30.33 ↑16.64	27.94 ↑7.43	19.23 ↑1.14	39.04 ↑10.74
	KD Sort	55.69 ↑6.54	29.31 ↑18.53	60.33 ↑2.72	32.44 ↑18.74	27.42 ↑6.90	9.55 ↓8.54	35.79 ↑7.48
	Active-Prompt	64.29 ↑15.15	30.03 ↑19.25	63.31 ↑5.70	29.61 ↑15.91	29.11 ↑8.60	13.74 ↓4.34	38.35 ↑10.04
	APIE	64.97 ↑15.82	32.61 ↑21.82	65.48 ↑7.87	32.57 ↑18.88	31.40 ↑10.89	14.11 ↓3.97	40.19 ↑11.88
Deepseek-R1-14B	ZSL	50.57 ↑0.00	17.76 ↑0.00	63.52 ↑0.00	27.30 ↑0.00	26.19 ↑0.00	11.06 ↑0.00	32.73 ↑0.00
	RSL	57.88 ↑7.30	31.50 ↑13.73	63.97 ↑0.46	27.83 ↑0.54	26.87 ↑0.68	11.30 ↑0.24	36.56 ↑3.83
	KD Sort	51.47 ↑0.90	31.24 ↑13.47	63.48 ↓0.04	33.11 ↑5.81	29.17 ↑2.98	10.75 ↓0.31	36.53 ↑3.80
	Active-Prompt	57.34 ↑6.77	41.19 ↑23.42	64.93 ↑1.41	32.13 ↑4.83	30.82 ↑4.63	11.48 ↑0.43	39.65 ↑6.91
	APIE	59.28 ↑8.71	41.95 ↑24.18	66.32 ↑2.81	34.62 ↑7.33	33.09 ↑6.90	13.58 ↑2.52	41.47 ↑8.74
Deepseek-v3-660B	ZSL	65.76 ↑0.00	24.87 ↑0.00	68.51 ↑0.00	31.74 ↑0.00	33.44 ↑0.00	20.30 ↑0.00	40.77 ↑0.00
	RSL	69.71 ↑3.95	33.31 ↑8.44	71.29 ↑2.79	42.32 ↑10.58	40.43 ↑6.99	20.35 ↑0.04	46.23 ↑5.46
	KD Sort	69.25 ↑3.49	38.62 ↑13.75	72.57 ↑4.06	51.01 ↑19.27	37.98 ↑4.54	14.20 ↓6.10	47.27 ↑6.50
	Active-Prompt	72.48 ↑6.72	47.33 ↑22.46	72.75 ↑4.25	47.66 ↑15.92	41.89 ↑8.46	18.90 ↓1.41	50.17 ↑9.40
	APIE	72.73 ↑6.97	48.56 ↑23.68	72.86 ↑4.36	50.20 ↑18.46	42.05 ↑8.61	21.45 ↑1.15	51.31 ↑10.54

Table 2: Overall performance (F1-score) comparison of APIE against baseline methods across four information extraction benchmarks and four LLM backbones. We report F1-scores for both Named Entity Recognition (NER) and, where applicable, Relation Extraction (RE). The best-performing method for each metric is **bold**, and the absolute improvement over the ZSL baseline is reported alongside the main score. APIE consistently reaches stable and highly competitive results.

optimal exemplar selection. Nonetheless, APIE maintains a consistent and measurable lead even at the largest scale (e.g., Deepseek-R1-14B), affirming the broad utility of our approach. Importantly, we observe that scaling alone does not resolve performance instability: as shown in Figure 3, random prompting continues to yield high variance even on high-capacity models, whereas APIE achieves remarkably stable results across all scales, underscoring its robustness and practicality in real-world deployments.

For RQ3: How are the different uncertainty signals distributed, and what does their behavior reveal about the nature of model confusion in IE tasks? We analyze the empirical distributions of the uncertainty scores from our experiments. We visualize these distributions using violin plots in Figure 5. The uncertainty signals are complementary and capture orthogonal aspects of difficulty in model behavior. The violin plots reveal that the three signals have distinct distributional characteristics. For instance, we observe cases where a sample has minimal format uncertainty ($U_f = 0.0$) but maximal content uncertainty ($U_c = 1.0$), indicating the model can generate the correct syntax but is confused about the semantic content. Conversely, other samples show high format uncertainty but low content uncertainty.

Analysis

Ablation Study. To dissect the contribution of each component in our framework, we conduct an ablation study using the DeepSeek-R1-14B model. As shown in Table 3, removing any single component—disagreement-based un-

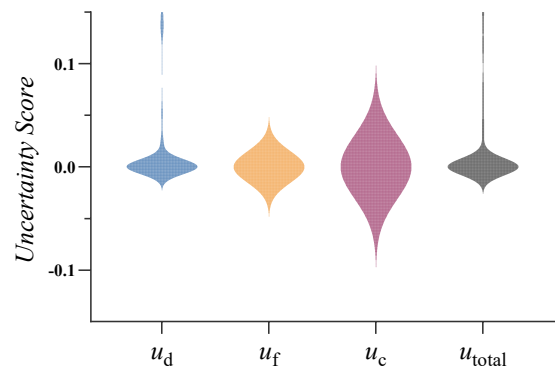


Figure 5: Uncertainty score distributions for U_d , U_f , and U_c on the ACE04-NER dataset, evaluated using the DeepSeek-R1-14B model.

certainty (U_d), format-level uncertainty (U_f), content-level uncertainty (U_c), or the pattern-guided prompt—leads to a noticeable performance drop across both NER and RE tasks. This indicates that each module contributes non-trivially to the overall performance of APIE. Notably, the exclusion of U_d causes the largest decline in F1-score, underscoring its critical role as a foundational signal of model uncertainty. The U_f and U_c metrics further complement this base signal, refining the selection toward structurally and semantically ambiguous samples. Additionally, the removal of pat-

Methods	ACE04-NER		CoNLL04 NER		CoNLL04 RE		Average	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Full APIE	35.04	41.95	55.71	66.32	32.46	34.62	41.07	47.63
w/o Pattern Prompt (p)	25.38	32.58	49.44	59.77	10.04	10.66	28.29	34.34
w/o Format Uncertainty (\mathcal{U}_f)	31.71	39.39	53.55	64.17	29.10	30.88	38.12	44.81
w/o Content Uncertainty (\mathcal{U}_c)	33.49	41.04	53.93	64.41	31.23	32.90	39.55	46.12
w/o Disagreement (\mathcal{U}_d)	26.56	32.19	52.61	62.86	30.13	31.98	36.43	42.34

Table 3: Ablation study on the effectiveness of APIE components across NER and RE tasks. Removing any component leads to performance degradation, with \mathcal{U}_d and pattern prompts being particularly critical.

Input Sentence:

On the 22nd, the sixteen NATO countries asked the armed forces of the Serbian people of Bosnia-Herzegovina to withdraw...

Ground Truth Entities:

{ { 'type': 'Organization', 'text': 'the sixteen NATO countries' }, { 'type': 'Location', 'text': 'Bosnia-Herzegovina' }, { 'type': 'Location', 'text': 'Gerlaridy city' }, { 'type': 'Group', 'text': 'Serbian' }, { 'type': 'Organization', 'text': 'NATO' }, { 'type': 'Organization', 'text': 'United Nations' } }

Gemma-3-12B: { { 'type': 'Organization', 'text': 'NATO' }, { 'type': 'Organization', 'text': 'United Nations' }, { 'type': 'Location', 'text': 'Bosnia-Herzegovina' }, { 'type': 'Organization', 'text': 'the Serbian army' } }

Qwen-2.5-14B: { { 'type': 'Organization', 'text': 'the sixteen NATO countries' }, { 'type': 'Organization', 'text': 'the armed forces of the Serbian people of Bosnia-Herzegovina' } }

Deepseek-R1-14B: { { 'type': 'Organization', 'text': 'sixteen NATO countries' }, { 'type': 'Location', 'text': 'Bosnia-Herzegovina' }, { 'type': 'Organization', 'text': 'United Nations humanitarian aid' } }

Uncertainty: $\mathcal{U}_d = 0.828$, $\mathcal{U}_f = 0.0$, $\mathcal{U}_c = 0.556$

Explanation: Despite consistent format ($\mathcal{U}_f = 0.0$), models diverge in semantic interpretation—some view 'sixteen NATO countries' as a geopolitical entity, others as an organization. APIE detects this content-level ambiguity ($\mathcal{U}_c = 0.556$) and prioritizes such samples for targeted refinement.

Figure 6: Case study illustrating content-level uncertainty despite consistent format understanding.

tern prompts yields the sharpest degradation in RE accuracy, suggesting that structure-aware prompt design plays an essential role in guiding the model toward consistent and schema-compliant outputs. These results validate the necessity of each module in constructing informative and stable prompts for reliable few-shot IE.

Case Study. To characterize the high-utility exemplars selected by APIE, we perform individual cases (Figure 6) and a qualitative analysis combining both aggregate trends (Figure 7). The analysis reveals a key distinction in selection strategies between our method and baselines. As shown in the radar plot, the Active-Prompt baseline, which relies on surface-level disagreement, tends to select for samples that are simply longer, denser, and cause high raw disagreement (\mathcal{U}_d). In contrast, APIE demonstrates a more targeted selec-

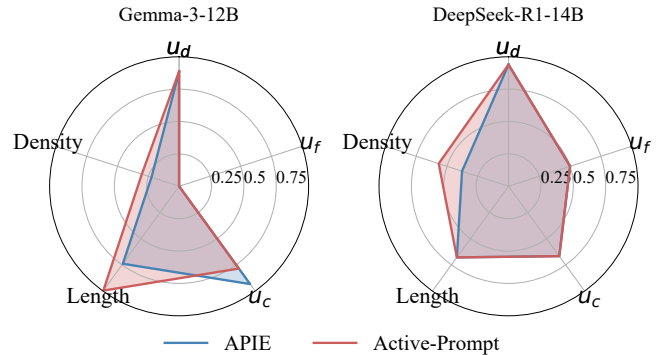


Figure 7: Comparison of exemplar characteristics selected by APIE and Active-Prompt on the ACE04-NER dataset.

tion strategy. While not always choosing the longest or densest samples, our framework shows a distinct and significant advantage in identifying instances with high content uncertainty (\mathcal{U}_c) and format uncertainty (\mathcal{U}_f). This suggests that our dual-level metric successfully moves beyond simple textual variance to pinpoint samples that are truly challenging in their semantic interpretation and structural requirements. The input sentence is syntactically straightforward, leading to zero format uncertainty ($\mathcal{U}_f = 0.0$). However, it elicits significant semantic confusion among different models regarding entity boundaries, resulting in high content uncertainty ($\mathcal{U}_c = 0.556$). By identifying such targeted, semantically ambiguous challenges, APIE enables more effective teaching for nuanced cases.

Conclusion and Outlook

In this work, we introduce APIE, a novel active prompting framework guided by introspective confusion to improve exemplar selection for few-shot information extraction. By disentangling model confusion into format and content uncertainty, APIE selects challenging exemplars to build more informative and robust prompts. Experiments show consistent gains over strong baselines, highlighting the value of structured uncertainty for improving LLM reliability. Future directions include more efficient uncertainty estimation, automatic hyperparameter tuning, and extensions to structured generation tasks like code generation, potentially combined with new diversity-based metrics.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 62506166, U2441285, 62222605), the Natural Science Foundation of Jiangsu Province (No. BK20251365), the China Postdoctoral Science Foundation (No. 2025M774283), the Yongjiang Talent Introduction Programme (No. 2021A-156-G) and the Ningbo Natural Science Foundation (No. 2024J020). This research is also sponsored by the DiDi GAIA Collaborative Research Funds (No. CCF-DiDi GAIA202507) and CAAI-MindSpore Open Fund (CAAIXSJLJJ 2025 MindSpore 01), developed on OpenI Community.

References

- Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.
- Chen, X.; Zhang, N.; Li, L.; Deng, S.; Tan, C.; Xu, C.; Huang, F.; Si, L.; and Chen, H. 2022a. Hybrid Transformer with Multi-level Fusion for Multimodal Knowledge Graph Completion. In Amigó, E.; Castells, P.; Gonzalo, J.; Carterette, B.; Culpepper, J. S.; and Kazai, G., eds., *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, 904–915. ACM.
- Chen, X.; Zhang, N.; Xie, X.; Deng, S.; Yao, Y.; Tan, C.; Huang, F.; Si, L.; and Chen, H. 2022b. Know-Prompt: Knowledge-aware Prompt-tuning with Synergistic Optimization for Relation Extraction. In Laforest, F.; Troncy, R.; Simperl, E.; Agarwal, D.; Gionis, A.; Herman, I.; and Médini, L., eds., *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, 2778–2788. ACM.
- DeepSeek-AI; Guo, D.; Yang, D.; Zhang, H.; and et al. 2025a. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.
- DeepSeek-AI; Liu, A.; Feng, B.; Xue, B.; and et al. 2025b. DeepSeek-V3 Technical Report. arXiv:2412.19437.
- Diao, S.; Wang, P.; Lin, Y.; Pan, R.; Liu, X.; and Zhang, T. 2024. Active prompting with chain-of-thought for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1330–1350.
- Gal, Y.; Islam, R.; and Ghahramani, Z. 2017. Deep bayesian active learning with image data. In *International conference on machine learning*, 1183–1192. PMLR.
- Han, R.; Yang, C.; Peng, T.; Tiwari, P.; Wan, X.; Liu, L.; and Wang, B. 2023. An empirical study on information extraction using large language models. *arXiv preprint arXiv:2305.14450*.
- Lewis, D. D.; and Gale, W. A. 1994. A Sequential Algorithm for Training Text Classifiers. In Croft, W. B.; and van Rijsbergen, C. J., eds., *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland, 3-6 July 1994 (Special Issue of the SIGIR Forum)*, 3–12. ACM/Springer.
- Liu, J.; Shen, D.; Zhang, Y.; Dolan, W. B.; Carin, L.; and Chen, W. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd workshop on knowledge extraction and integration for deep learning architectures*, 100–114.
- Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM computing surveys*, 55(9): 1–35.
- Lu, Y.; Liu, Q.; Dai, D.; Xiao, X.; Lin, H.; Han, X.; Sun, L.; and Wu, H. 2022. Unified Structure Generation for Universal Information Extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5755–5772.
- Luan, Y.; He, L.; Ostendorf, M.; and Hajishirzi, H. 2018. Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3219–3232. Brussels, Belgium: Association for Computational Linguistics.
- Ma, Y.; Cao, Y.; Hong, Y.; and Sun, A. 2023. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 10572–10601.
- Margatina, K.; et al. 2023. Active Learning Principles for In-Context Learning with Large Language Models. In *arXiv preprint arXiv:2305.14264*.
- Mitchell, A.; Strassel, S.; Huang, S.; and Zakhary, R. 2005. ACE 2004 Multilingual Training Corpus.
- Pai, L.; Gao, W.; Dong, W.; Ai, L.; Gong, Z.; Huang, S.; Zongsheng, L.; Hoque, E.; Hirschberg, J.; and Zhang, Y. 2024. A survey on open information extraction from rule-based model to large language model. *Findings of the association for computational linguistics: EMNLP 2024*, 9586–9608.
- Qian, K.; Sang, Y.; Bayat, F.; Belyi, A.; Chu, X.; Govind, Y.; Khorshidi, S.; Khot, R.; Luna, K.; Nikfarjam, A.; et al. 2024. APE: Active learning-based tooling for finding informative few-shot examples for LLM-based entity matching. In *Proceedings of the Fifth Workshop on Data Science with Human-in-the-Loop (DaSH 2024)*, 1–3.
- Qwen; Yang, A.; Yang, B.; Zhang, B.; and et al. 2025. Qwen2.5 Technical Report. arXiv:2412.15115.
- Roth, D.; and Yih, W.-t. 2004. A Linear Programming Formulation for Global Inference in Natural Language Tasks. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*, 1–8. Boston, Massachusetts, USA: Association for Computational Linguistics.
- Settles, B. 2009. Active learning literature survey. *University of Wisconsin-Madison Department of Computer Sciences*.
- Team, G.; Kamath, A.; Ferret, J.; and et al. 2025. Gemma 3 Technical Report. arXiv:2503.19786.

Tjong Kim Sang, E. F.; and De Meulder, F. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In Daelemans, W.; and Osborne, M., eds., *Proceedings of CoNLL-2003*, 142–147. Edmonton, Canada.

Wang, S.; Sun, X.; Li, X.; Ouyang, R.; Wu, F.; Zhang, T.; Li, J.; Wang, G.; and Guo, C. 2025. Gpt-ner: Named entity recognition via large language models. In *Findings of the association for computational linguistics: NAACL 2025*, 4257–4275.

Xie, T.; Wu, C. H.; Shi, P.; Zhong, R.; Scholak, T.; Yasunaga, M.; Wu, C.; Zhong, M.; Yin, P.; Wang, S. I.; Zhong, V.; Wang, B.; Li, C.; Boyle, C.; Ni, A.; Yao, Z.; Radev, D.; Xiong, C.; Kong, L.; Zhang, R.; Smith, N. A.; Zettlemoyer, L.; and Yu, T. 2022. UnifiedSKG: Unifying and Multi-Tasking Structured Knowledge Grounding with Text-to-Text Language Models. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, 602–631. Association for Computational Linguistics.

Zhang, J.; You, Z.; Wang, J.; and Le, X. 2025. SAIL: Sample-Centric In-Context Learning for Document Information Extraction. In Walsh, T.; Shah, J.; and Kolter, Z., eds., *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, 25868–25876. AAAI Press.