An LLM-based Framework for Domain-Specific Information Extraction: A Case Study in Computer Science and Chemistry

Xungang Gu[1] **and Yangjie Tian**[2] **and Ning Li**[3] **and Meng Liu**[3] **and Ruohua Xu**[3]

He Zhang[3*] **and Hanqiu Liu**[4] **and Yongpan Sheng**[5] **and Ming Liu**[1]

[1]School of Information Technology, Deakin University, Melbourne, VIC 3125, Australia
[2]Institute for Sustainable Industries and Liveable Cities, Victoria University, Melbourne, VIC 3011, Australia
[3]Kexin Technology, Beijing 100012, China
[4]Monash Business School, Monash University, Caulfield East, VIC 3145, Australia
[5]College of Computer and Information Science, Southwest University, Chongqing 400715, China
x.gu@deakin.edu.au; yangjie_tian@163.com; {lining, liumeng, xuruohua, zhanghe}@kxsz.net
hanqiuliu2@gmail.com; shengyp2011@gmail.com; m.liu@deakin.edu.au

Abstract

Information extraction (IE) in specialized domains like computer science and chemistry is challenged by the poor generalization of traditional models and the knowledge deficits of general-purpose Large Language Models (LLMs). We introduce a robust, LLM-based framework featuring two core contributions: an end-to-end training and inference paradigm that combines continual pre-training (CPT) for knowledge injection, supervised fine-tuning (SFT) for task alignment, and retrieval-augmented generation (RAG) for inference-time enhancement; and a novel LLM-assisted data annotation pipeline for the efficient creation of high-quality training data. Comprehensive experiments demonstrate that while fine-tuning alone yields strong in-domain performance, our complete framework exhibits superior robustness and generalization. It consistently achieves state-of-the-art results in challenging domain-shift and novel-schema scenarios, validating our integrated approach for building adaptable and high-performance domain-specific IE systems.

1 Introduction

Domain-specific information extraction is crucial for converting unstructured data, such as scientific text or chemical descriptions, into structured knowledge, which in turn enables downstream tasks like knowledge graph construction and scientific discovery (Dagdelen et al., 2024). However, traditional IE models suffer from poor generalization; models trained for a specific schema often fail to generalize to new entity and relation types or different data domains (Peng et al., 2021).

Large Language Models (LLMs) offer a promising alternative: they exhibit strong generalization and can follow instructions to extract information without task-specific architecture. However, their accuracy on specialized IE tasks often fails to surpass traditional, domain-trained models (Han et al., 2023), as general-purpose LLMs lack detailed domain knowledge (e.g., chemical nomenclature (Castro Nascimento and Pimentel, 2023), scientific terminology) and are not optimized for the nuances of structured extraction (Dagdelen et al., 2024). To harness the generalization capabilities of LLMs for high-quality, domain-specific extraction while overcoming their inherent limitations, several key challenges must be addressed:

- **Injecting domain knowledge:** The LLM must be enriched with specialist knowledge of target domains (here, computer science and chemistry) to overcome its knowledge gaps.
- **Improving IE capabilities:** The LLM should be fine-tuned to adeptly perform the information extraction task itself, enabling it to accurately identify and structure entities and relations from complex texts, thereby surpassing the performance of traditional IE models.
- **Enhancing cross-schema generalization:** The solution should handle different entity–relation type schemas and adapt to new types or distributions with minimal re-training, leveraging the LLM's generalization.

To address these challenges, we design and implement a comprehensive, LLM-based information extraction framework that significantly enhances a model's domain knowledge, extraction accuracy, and generalization stability. Our core contributions include:

- An end-to-end training and inference paradigm that integrates **Continual Pre-**

* Corresponding author.

Proceedings of The 23rd Annual Workshop of the Australasian Language Technology Association, pages 101–111

training (CPT), **Supervised Fine-tuning (SFT)**, and **Retrieval-Augmented Generation (RAG)**. This paradigm is designed to systematically inject domain knowledge, align the model with the specific extraction task, and leverage external knowledge to enhance its ability to handle complex cases.

- A novel **LLM-assisted data annotation pipeline** that efficiently constructs high-quality, domain-specific training datasets at a low cost. By leveraging multi-model collaboration, consensus fusion, and a reward model gating mechanism, this pipeline effectively mitigates the data bottleneck problem.

We validate our framework through a comprehensive evaluation in the computer science and chemistry domains across three rigorous scenarios: in-domain, domain-shift, and novel-schema settings. Our experimental results demonstrate that while fine-tuning alone yields strong in-domain performance, our complete framework exhibits superior robustness and generalization. This advantage becomes particularly evident when faced with data distribution shifts and unseen schemas, where our full framework outperformed the fine-tuning-only baseline by a margin of 3.0 and up to 6.7 entity F1 points, respectively. This research charts a clear and effective path toward building high-performance, domain-specific IE systems that can adapt to the variable conditions of real-world applications.

2 The Proposed Framework

Our approach transforms a general-purpose LLM into a specialized and robust information extractor through an integrated, three-stage workflow, as illustrated in Figure 1. Below is the detailed description of each stages.

2.1 Continual Pre-training for Domain Knowledge

To inject domain-specific knowledge into the base LLM, we employ Continual Pre-training. This foundational stage adapts the model's language understanding to the target domains by exposing it to a large corpus of specialized text. Our approach is carefully designed not only to acquire new knowledge but also to mitigate the catastrophic forgetting of the model's general capabilities.

2.1.1 Objective and Forgetting Mitigation

The primary objective of CPT is to enrich the LLM with the terminology, concepts, and stylistic nuances of the target domains. A key challenge in this process is mitigating catastrophic forgetting (Gu et al., 2024). To address this, we curate a balanced mixture of general and domain-specific data. Rather than relying on exhaustive grid searches, we follow the D-CPT scaling law (Que et al., 2024) to determine the optimal domain/general data mix. The D-CPT is achieved by modeling the monotonic trade-off between domain loss and general-corpus loss from a small set of pilot runs, allowing us to select a domain data share r_d that minimizes domain validation loss while constraining the increase of general validation loss within a predefined tolerance budget T.

2.1.2 Corpus Curation and Cleaning

We assemble the domain corpora from established scientific sources, including arXiv CS categories and the ACL Anthology for computer science, and PubMed/PMC subfields and patent corpora for chemistry. Prior to CPT, we execute a deterministic cleaning pipeline to ensure data quality. This pipeline comprises: (i) language identification and basic normalization; (ii) de-boilerplating and removal of non-content sections; (iii) OCR and markup error repair; (iv) topicality filtering using domain-specific lexicons; and (v) exact and near-duplicate removal via shingled MinHash/LSH. This process yields a clean, diverse, and on-topic corpus for both domains.

2.2 Dataset Construction

A high-quality, comprehensive training dataset for information extraction (IE) is the cornerstone of our approach. We construct it by (i) gathering, cleaning, and unifying existing open IE datasets in our target domains (computer science and chemistry) and (ii) creating a custom, LLM-assisted annotated dataset tailored to our schemas.

2.2.1 Gathering and Comparing Open IE Datasets

We first collect relevant public IE datasets, then standardize and merge them. Below summarizes the key resources for computer science and chemistry/biomedical domains, respectively.

Computer Science. We include **SciERC** (Luan et al., 2018) with six entity types (Task, Method, Metric, Material, OtherScientificTerm, Generic)

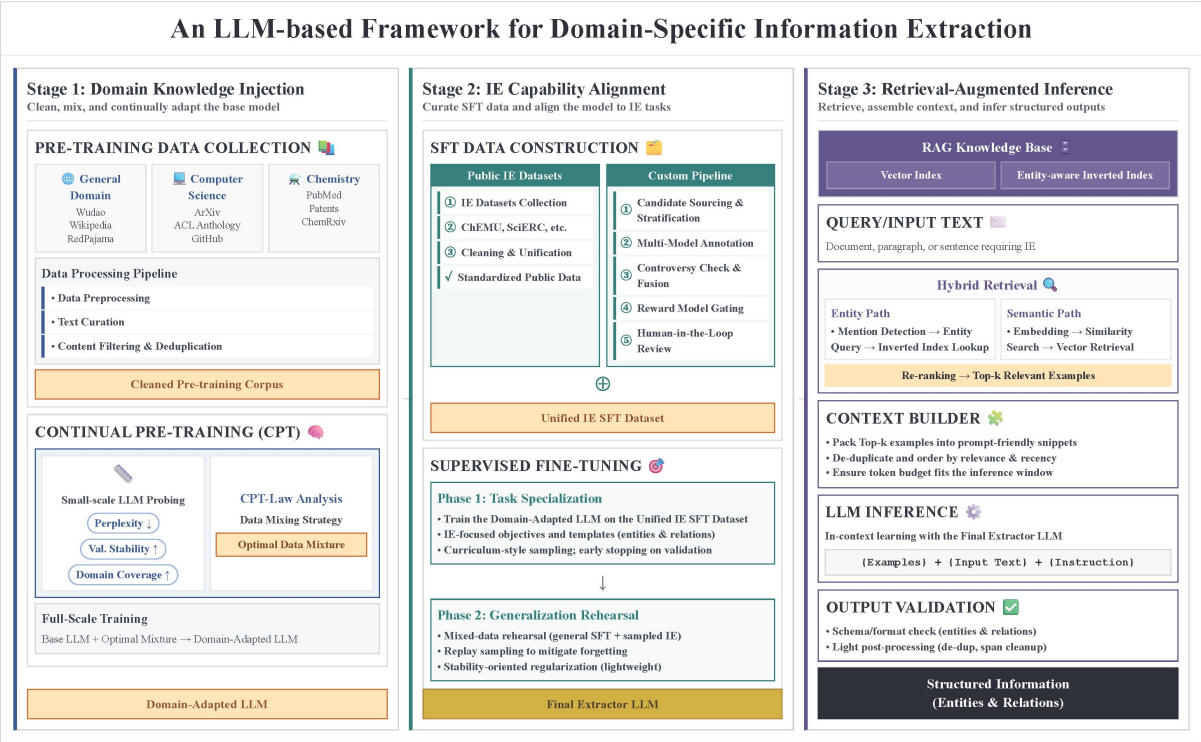

Figure 1: LLM-based Framework for Domain-Specific Information Extraction.

and seven relation types (e.g., *Used-for*, *Compare*, *Part-of*). **SciREX** (Jain et al., 2020) contributes cross-sentence coreference and document-level relations. **SciER** (Zhang et al., 2024) adds a large number of entities/relations and focuses on Dataset, Method, and Task, with fine-grained relations and an out-of-distribution split. To increase diversity, we incorporate **CrossNER** (Peng et al., 2021) (a cross-domain NER collection; we use the AI/Science splits) and **CrossRE** (Wang et al., 2022) (a cross-domain relation extraction dataset with multi-label relations). This multi-source integration spans multiple schemas and helps models handle schema heterogeneity.

Chemistry/Biomedical. We compile datasets capturing chemical/biomedical entities and relations. **NLM-Chem** (Kim et al., 2021) provides rich full-text chemical NER. For relations, **ChemProt** (Krallinger et al., 2017) offers sentence-level relation labels (e.g., inhibitor, upregulator). **BioRED** (Luo et al., 2022a) includes multiple entity types and document-level relations, marking novelty. From chemical patents, **ChEMU-2020** (Nguyen et al., 2020) targets reaction extraction with entities like Reactant, Product, Catalyst, Solvent, and conditions (Temperature, Time), plus event-style relations. **EnzChemRED** (Lai et al., 2024) focuses on enzyme–reaction relations, linking to ontologies (e.g., ChEBI, UniProt). Together these resources align with our chemistry use-cases and add complementary schemas.

Data Cleaning and Unification. To unify the diverse datasets, we standardize all annotations into a consistent JSON format and resolve notational conflicts. To manage the inherent schema heterogeneity, each training prompt explicitly defines the target entity and relation types for the given instance. This unification process is designed to expand data coverage and enhance the model's robustness to schematic variations.

2.2.2 LLM-Assisted Custom Data Annotation

To supplement the public corpora, we designed and implemented a novel LLM-assisted annotation pipeline to efficiently create high-quality, schema-specific training data. Our approach systematically reduces manual effort and ensures data quality through a multi-stage workflow. As show in Figure 2, this process involves: (1) strategic candidate sourcing, (2) parallel labeling by multiple LLMs to generate diverse annotations, (3) agreement-based fusion to consolidate results, (4) quality control via a calibrated reward model, and (5) a focused human-in-the-loop process for arbitration and feedback. Ultimately, this pipeline provides a cost-

effective methodology for generating high-quality, tailored training data, ensuring both label precision and broad data coverage. A detailed description of each stage is provided in Appendix A.

2.3 Model Fine-tuning and Training Strategies

Following domain-adaptive CPT, we perform SFT to align the model with the specific task of structured information extraction. This stage uses a carefully composed mixture of our curated, chat-formatted IE dataset and a high-quality general instruction-following dataset. This approach is designed to teach the model how to generate accurate and well-formed entity-relation structures, while simultaneously managing the trade-off between domain specialization and its broader instruction-following capabilities.

2.3.1 Training Variants

To systematically evaluate the impact of different base models, training stages, we designed two distinct experimental cases. The CPT dataset consists of a mix of general and domain-specific texts, while the SFT dataset combines our constructed domain-specific IE data with general instruction-following examples.

- **Case 1 (SFT-only):** A general-purpose chat model is fully fine-tuned on the SFT dataset. This evaluates the performance ceiling of a standard chat model without domain-specific pre-training.
- **Case 2 (CPT-SFT):** The base model first undergoes full-parameter CPT with the CPT dataset, followed by full-parameter SFT on the SFT dataset. This case represents our full, two-stage proposal.

2.3.2 Data Composition and Mixture Strategy

Balancing task-specific specialization with general capabilities in SFT can trigger training conflicts or catastrophic forgetting (Dong et al., 2023). We therefore adopt a two-stage curriculum with mixture optimization. In Stage 1, the model is fine-tuned exclusively on our IE dataset to acquire structured extraction skills. In Stage 2, training continues on a mixture of general SFT data plus a small sampled subset of the IE data; the general data restores broad abilities while the sampled IE acts as rehearsal to mitigate forgetting. To determine the optimal data composition for this second stage, we took inspiration from Gu et al.(Gu et al., 2025) to set initial candidate ratios (e.g., 300:1, 150:1 of general-to-IE data) and then identified the best-performing mixture through small-scale empirical tests. This two-stage curriculum effectively resolves the training conflict, allowing us to maximize the model's IE performance without a significant loss of its general instruction-following abilities.

2.4 Retrieval-Augmented Generation (RAG) for IE

Even after fine-tuning, a model may struggle with complex cases or long-tail knowledge (Liao et al., 2024). We therefore integrate a RAG component that equips the extractor with an explicit, query-time knowledge base (KB) of reference examples and facts. The core idea is to maintain a repository of labeled examples so that, when extracting from new inputs, the model can draw on similar past cases to assist its predictions.

2.4.1 Knowledge Base Construction

We populate the KB with high-quality extraction examples, drawn from two primary sources:

- **Confidently labeled data**: This includes high-quality, LLM-annotated samples and ground-truth data from public datasets, consistent with the data construction methodology detailed in Sections 2.2.1 and 2.2.2. Concretely, we retain the top 10–20% most confident instances per batch—measured by the reward-model score after JSON-schema and span-alignment validation—and we explicitly exclude any instance used for supervised fine-tuning from the KB.
- **Manually verified cases**: Difficult examples that the pipeline initially withheld due to high model disagreement (e.g., high controversy; see Appendix A) or validation issues, and were subsequently reviewed by domain experts (two annotators with a third arbitrator). We also include failure cases surfaced in downstream use that were corrected by experts. These curated items provide valuable exemplars for resolving ambiguity and expanding coverage of hard cases.

Each KB entry is stored as a *(text, extraction)* pair. To support efficient look-up, we index the KB using a hybrid approach that combines semantic and

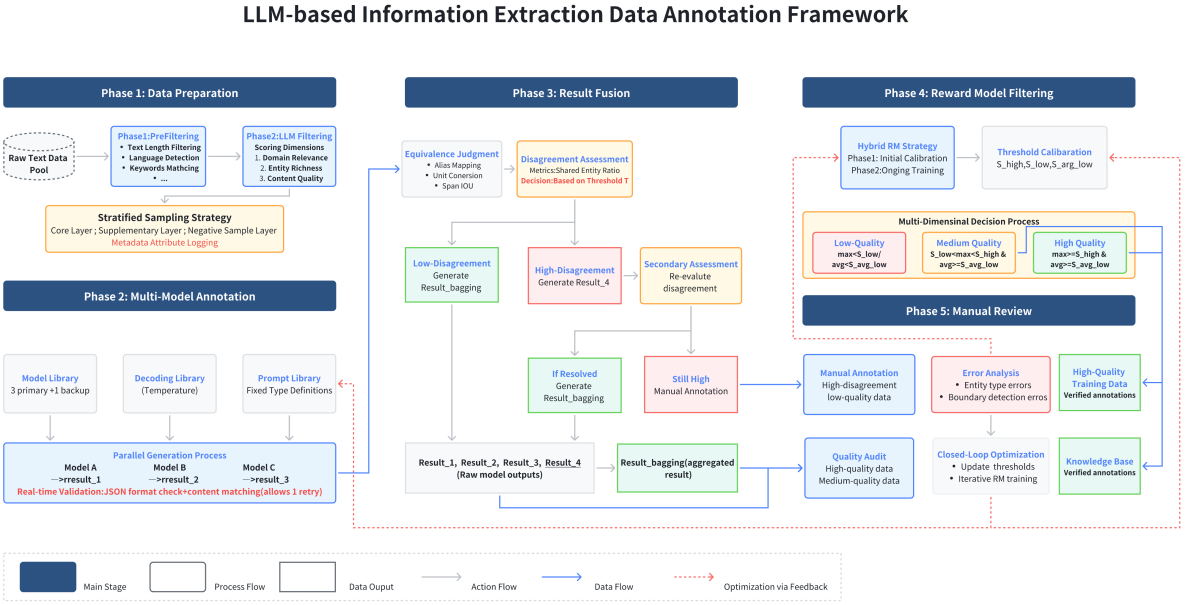

Figure 2: LLM-assisted custom data annotation workflow for IE.

lexical matching: a vector index (FAISS) is built on text embeddings for semantic similarity, and an entity-aware inverted index is used to rapidly match entries containing shared entity mentions.

2.4.2 Two-Stage Retrieval

Given a new input passage, we employ a two-stage retrieval process to gather relevant exemplars from the knowledge base.

(1) Entity-based Retrieval. First, a fine-tuned BERT-based model for entity mention detection extracts potential entity strings from the input text, without performing classification. These identified entity strings are then used to query the entity-aware inverted index, retrieving all KB entries that contain one or more of these exact entities. This stage ensures high topical relevance between the retrieved examples and the input.

(2) Semantic Retrieval. Concurrently, we use a Sentence-BERT model to generate a dense vector embedding for the entire input passage. This embedding is used to perform a nearest-neighbor search against the vector index of the KB, retrieving passages that are contextually and semantically similar, even if they lack shared entities.

Result Combination and Re-ranking The top-k results from both retrieval methods are aggregated to form a candidate pool. We then re-rank these candidates using a hybrid score that balances lexical matching (via entities) with semantic similarity.

To ensure both components are on a comparable scale, we first normalize the entity overlap into a score, $\text{sim}_{\text{ent}}(e)$. This score is defined as the fraction of query entities (E_q) found in the candidate document's entities (E_e):

$$\text{sim}_{\text{ent}}(e) = \frac{|E_q \cap E_e|}{|E_q|}$$

The final re-ranking score is then a weighted linear combination of this entity score and the semantic similarity score, $\text{sim}_{\text{sem}}(e)$:

$$\text{score}(e) = \lambda \cdot \text{sim}_{\text{ent}}(e) + (1 - \lambda) \cdot \text{sim}_{\text{sem}}(e)$$

Here, $\lambda \in [0, 1]$ is a tunable hyperparameter that controls the relative importance of entity-based relevance versus overall contextual similarity. Its optimal value is determined based on performance on a held-out validation set.

2.4.3 Prompt Integration and Dynamic Improvement

The top-ranked retrieved `(text, extraction)` pairs are formatted as in-context examples to guide the model's generation. Crucially, to prevent schema confusion, retrieved candidates are first filtered to ensure they match the target schema of the current task. These schema-consistent examples provide the model with on-the-fly guidance on output format and extraction logic for challenging cases. To ensure quality and prevent simple copying, we apply several safeguards: filtering out

examples with high lexical overlap to the input, using minimal text snippets to maintain a concise prompt, and instructing the model to use the examples for reference only.

Furthermore, we implement a 'data flywheel' to ensure the RAG system's long-term effectiveness. High-confidence new extractions, often validated through a lightweight human verification step, are continuously added back into the knowledge base. This iterative process progressively enriches the KB in both scale and quality, enhancing future RAG performance without the need for frequent model retraining.

3 Experiments

Our experiments are designed to rigorously evaluate our proposed framework and validate the contributions of its core components. We structure our evaluation around three key research questions:

- **RQ1: Overall Performance.** How does our full framework perform against baselines(general purpose LLM) on domain-specific IE tasks, under both in-domain and domain-shifted conditions?
- **RQ2: Ablation Study.** What is the individual contribution of each key component(CPT and RAG) to the final extraction performance?
- **RQ3: Generalization Analysis.** How effectively does our framework generalize to novel, unseen schemas, demonstrating its adaptability?

3.1 Experimental Setup

3.1.1 Training and Test Data

Training Data. Our models for the computer science (CS) and chemistry domains were trained separately. For the **CPT** stage, we started with 500k general-domain texts and 100k texts for each specific domain. Following the data mixing search strategy from Section 2.1, the final compositions were a 400k:100k ratio of general-to-CS data for the computer science model, and a 350k:100k ratio for the chemistry model. For the **SFT** stage, each domain's dataset comprised 8,000 high-quality IE instances (5,000 from public datasets and 3,000 constructed via our pipeline). Based on the strategy in Section 2.3.2, each domain model was first fine-tuned on the complete set of 8,000 IE instances for specialization. The second stage then employed a mixed dataset for generalization and rehearsal, consisting of 50k general and 1k sampled CS instances for computer science, and 100k general and 1k sampled chemistry instances for chemistry.

Test Data. We evaluate all models on three distinct test sets, each containing 500 instances for computer science and 500 for chemistry:

- **Test Set A (IID):** This set was constructed using our in-house annotation pipeline and has a data distribution similar to our self-built training data. It measures the model's core extraction accuracy on a familiar data distribution.
- **Test Set B (Domain Shift):** Also constructed in-house, this set features a noticeable domain shift. For instance, while the CS data in Test Set A focuses on the AI subfield, Test Set B contains texts from non-AI subfields, sourced using specific keywords and categories. This set assesses model robustness.
- **Test Set C (Novel Schema):** This set uses the official test splits of public datasets—**SciERC** for computer science and **ChemU** for chemistry. To ensure a fair test of schema generalization, the training sets of these two datasets were completely excluded from our model training while keeping in RAG knowledge base.

3.1.2 Models for Comparison

All our trained models are based on the **Qwen2.5-7B** large language model. We compare the following configurations:

- **Chat-only:** The publicly available **Qwen2.5-7B-Chat** model, used directly without any fine-tuning, serves as a strong baseline.
- **SFT-only:** This model is initialized from **Qwen2.5-7B-Chat** and then fully fine-tuned on our SFT datasets for each domain.
- **CPT-SFT:** Our proposed model without retrieval. It starts from the **Qwen2.5-7B-Base** model, first undergoes CPT with our mixed-domain corpora, and is then SFT.
- **RAG Models:** We also evaluate three RAG-enhanced variants: **Chat-RAG**, **SFT-RAG** and our full system, **CPT-SFT-RAG**, to measure the impact of retrieval.

3.2 Main Results and Analysis

We evaluate entity and relation extraction performance using F1-score. Table 1 presents the comprehensive evaluation of all model variants across the six test sets for both Computer Science (CS) and Chemistry (Chem) domains. Our analysis is structured around the three research questions to dissect these results.

3.2.1 RQ1: Overall Performance

As shown in Table 1, the baseline 'Chat-only' model exhibits modest performance, confirming that general-purpose LLMs struggle with the structured and specialized nature of domain-specific IE tasks. The introduction of fine-tuning ('SFT-only') provides a substantial performance leap, highlighting the necessity of task-specific adaptation.

Our full proposed model, CPT-SFT-RAG, demonstrates the most robust overall performance. While SFT-RAG achieves the highest scores on the in-domain Test Set A for both CS and Chemistry, CPT-SFT-RAG excels under more challenging conditions. It secures the top F1-scores on Test Set B (Domain Shift) and Test Set C (Novel Schema) across both domains. This superior performance under distribution shifts and on unseen schemas validates our framework's primary goal: to create an IE system that not only performs well but also generalizes robustly, effectively mitigating the brittleness of traditional models and the knowledge deficit of general LLMs.

3.2.2 RQ2: Ablation Study

By comparing different model configurations, we can isolate the contributions of CPT and RAG.

Impact of Continual Pre-training (CPT). A comparison between SFT-only and CPT-SFT reveals the crucial role of domain knowledge injection. In the CS domain on Test A, SFT-only slightly outperforms CPT-SFT. This may be attributed to the base LLM's existing familiarity with computer science concepts, where direct SFT on a chat-tuned model can be highly effective. However, in the Chemistry domain, which contains more specialized terminology and symbolic representations, CPT-SFT surpasses SFT-only on Test A. This suggests that the greater the knowledge gap between the general domain and the target domain, the more significant the benefit of CPT.

Furthermore, across both domains on the more challenging Test B and Test C, CPT-SFT consistently outperforms SFT-only. This demonstrates that CPT provides a stronger and more generalizable knowledge foundation, enhancing the model's stability against data distribution shifts and its adaptability to new schemas.

Impact of Retrieval-Augmented Generation (RAG). RAG consistently improves performance across all base models. The most dramatic gain is seen when applying it to the baseline, where Chat-RAG significantly outperforms Chat-only. However, the performance uplift from RAG diminishes as the base model becomes more capable (i.e., the gain from SFT-only to SFT-RAG is larger than from CPT-SFT to CPT-SFT-RAG). This indicates that while RAG is a powerful tool, its marginal benefit is related to the base model's inherent instruction-following and domain understanding capabilities.

Interestingly, in the knowledge-intensive Chemistry domain, the improvements from CPT are often more pronounced than those from RAG. For instance, CPT-SFT achieves a higher entity F1 across all three Chemistry test sets compared to SFT-RAG, despite SFT-RAG having access to in-context examples. This suggests that for highly specialized domains, knowledge internalized through CPT provides a more robust and fundamental capability than knowledge supplied externally via retrieval at inference time.

3.2.3 RQ3: Generalization to Novel Schemas

The results on Test Set C (SciERC for CS and ChEMU for Chem) directly measure the models' ability to generalize to unseen schemas defined only in the prompt. The CPT-SFT-RAG model achieves the highest performance on both entity and relation extraction in this setting. On SciERC, it reaches a 69.9 entity F1, and on the complex ChEMU dataset, it achieves a 84.9 entity F1, leading all other configurations.

This success highlights the synergy of our framework's components. CPT provides the model with a deep understanding of the domain's entities and their typical interactions. SFT fine-tunes its ability to follow structured instructions. Finally, RAG provides concrete examples of the novel schema, guiding the model to apply its knowledge in the required format. The combination of internalized domain knowledge and in-context schema exemplars allows CPT-SFT-RAG to adapt more effectively than models relying on only one of these aspects.

Table 1: Overall performance (F1-score) of all model variants on Test Set A (IID), Test Set B (Domain Shift), and Test Set C (Novel Schema). Best results in each column are in **bold**, and other notable results are underlined.

Case	Computer Science (CS)						Chemistry (Chem)					
	Test A		Test B		SciERC		Test A		Test B		ChEMU	
	Entity	Relation	Entity	Relation	Entity	Relation	Entity	Relation	Entity	Relation	Entity	Relation
Chat-only	61.1	37.8	60.2	40.7	52.8	32.1	55.5	34.9	58.1	39.9	63.2	-
Chat-RAG	68.9	50.7	69.1	49.9	59.0	36.9	64.8	41.2	66.4	45.4	70.1	-
SFT-only	85.8	69.3	80.9	67.5	63.2	43.7	84.1	52.7	79.3	48.8	80.5	-
CPT-SFT	84.6	67.8	81.3	69.8	65.7	45.0	85.9	54.4	82.8	51.1	83.1	-
SFT-RAG	**86.9**	**75.6**	83.4	73.8	68.0	49.9	**86.8**	**57.3**	82.1	52.9	82.6	-
CPT-SFT-RAG	85.1	74.2	**83.9**	**74.1**	**69.9**	**51.8**	85.1	55.9	**84.2**	**54.0**	**84.9**	-

Summary of findings. Our experiments confirm that supervised fine-tuning is essential for high performance in IE tasks. However, to achieve robustness and generalization, further steps are necessary. Continual pre-training (CPT) is critical for building a solid domain knowledge foundation, especially in fields distant from the LLM's general training data. Retrieval-augmented generation (RAG) provides a consistent performance boost, particularly for weaker base models, by supplying relevant context at inference. The integrated CPT-SFT-RAG framework proves to be the most effective, achieving state-of-the-art performance on domain-shifted and novel-schema tasks, thereby demonstrating a practical path toward building powerful and adaptable domain-specific IE systems.

4 Related Work

Traditional Information Extraction (IE) models often fail to generalize across new domains and schemas, a limitation highlighted by benchmarks like CrossRE (Wang et al., 2022). To address this, research has pursued unified frameworks capable of handling heterogeneous structures, such as UIE and USM (Lu et al., 2022; Lou et al., 2023), and has leveraged multi-task instruction tuning to improve transfer learning (Wang et al., 2023). Another approach involves creating generalist zero-shot models like GLiNER and UniversalNER for broader entity coverage (Zaratiana et al., 2024; Zhou et al., 2023).

The advent of Large Language Models (LLMs) offers a new paradigm for IE, though studies show that off-the-shelf models can be inconsistent and often lag behind specialized systems (Han et al., 2023; Ma et al., 2023). Two primary strategies have emerged to enhance LLM-based IE. The first is **domain adaptation**, where models like BioGPT are fine-tuned on specialized corpora to internalize domain knowledge, leading to significant performance gains (Luo et al., 2022b; Dagdelen et al., 2024). The second is **retrieval augmentation**, which injects external evidence at inference time to improve accuracy and robustness, as demonstrated by frameworks like RUIE and RAMIE (Liao et al., 2024; Zhan et al., 2025). Our work integrates these successful strategies, combining domain-adaptive pre-training, supervised instruction tuning, and retrieval to build a robust and schema-flexible IE system.

5 Conclusion

We proposed an LLM-based framework for domain-specific Information Extraction. Experimental results in computer science and chemistry validate this integrated strategy. Our full CPT-SFT-RAG model demonstrates superior robustness and generalization, excelling in challenging domain-shift and novel-schema scenarios. Ablation studies highlight CPT's critical role in establishing a foundational domain understanding, especially for knowledge-intensive fields, while RAG provides a reliable performance boost. We find, however, that knowledge internalized via CPT is more fundamental to out-of-distribution generalization than examples supplied by RAG at inference time.

In summary, this work contributes an effective and generalizable methodology for developing high-performance, adaptable IE systems. Our framework offers a practical solution that balances accuracy with robustness, and our data annotation pipeline helps mitigate the critical data bottleneck, facilitating wider application and innovation in specialized scientific domains.

Limitations

Our framework's performance is benchmarked on a 7B-scale model, and its full-parameter training

is computationally intensive, which may limit scalability. While larger base models are expected to narrow in-domain gaps for SFT-only, we anticipate that CPT will remain most beneficial under domain shift and novel-schema settings, and that RAG's marginal utility will concentrate on tail errors and schema formatting rather than average-case gains. The RAG component's effectiveness is contingent on retrieval quality and introduces inference latency. Furthermore, our validation is currently confined to entity and relation extraction. Future work will explore scaling to larger models using parameter-efficient fine-tuning (PEFT), improving the retrieval mechanism, and extending our methodology to more complex IE tasks like event extraction, along with re-running our ablations at larger scales to empirically verify these expectations.

References

Camila M. Castro Nascimento and Alexandre S. Pimentel. 2023. Do large language models understand chemistry? a conversation with ChatGPT. *Journal of Chemical Information and Modeling*, 63(6):1649–1655.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S. Rosen, Gerbrand Ceder, Kristin A. Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1):1418.

G. Dong, H. Yuan, K. Lu, and 1 others. 2023. How abilities in large language models are affected by supervised fine-tuning data composition. *Computing Research Repository*, arXiv:2310.05492.

J. Gu, Z. Yang, C. Ding, and 1 others. 2024. CMR scaling law: Predicting critical mixture ratios for continual pre-training of language models. *Computing Research Repository*, arXiv:2407.17467.

X. Gu, M. Wang, Y. Tian, and 1 others. 2025. A comprehensive approach to instruction tuning for Qwen2.5: Data selection, domain interaction, and training protocols. *Computers*, 14(7):264.

R. Han, T. Peng, C. Yang, and 1 others. 2023. Is information extraction solved by ChatGPT? an analysis of performance, evaluation criteria, robustness and errors. *Computing Research Repository*, arXiv:2305.14450.

Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. Scirex: A challenge dataset for document-level information extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Sun Kim, Chih-Hsuan Wei, Robert Leaman, and Zhiyong Lu. 2021. Nlm-chem: a new resource for chemical entity recognition in pubmed full-text articles. *Scientific Data*, 8(1):205.

Martin Krallinger, Obdulia Rabal, Saber A. Akhondi, and et al. 2017. Overview of the BioCreative VI chemical–protein interaction track. In *Proceedings of the BioCreative VI Workshop*, Bethesda, MD. Describes the ChemProt corpus.

Po-Ting Lai, Elisabeth Coudert, Lucila Aimo, Kristian Axelsen, Lionel Breuza, Edouard de Castro, Marc Feuermann, Anne Morgat, Lucille Pourcel, Ivo Pedruzzi, Sylvain Poux, Nicole Redaschi, Catherine Rivoire, Anastasia Sveshnikova, Chih-Hsuan Wei, Robert Leaman, Ling Luo, Zhiyong Lu, and Alan Bridge. 2024. Enzchemred, a rich enzyme chemistry relation extraction dataset. *arXiv preprint arXiv:2404.14209*.

Xincheng Liao, Junwen Duan, Yixi Huang, and Jianxin Wang. 2024. Ruie: Retrieval-based unified information extraction using large language model. *Computing Research Repository*, arXiv:2409.11673.

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 Conference on Artificial Intelligence*, volume 37, pages 13318–13326.

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 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5755–5772, Dublin, Ireland. Association for Computational Linguistics.

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 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N. Arighi, and Zhiyong Lu. 2022a. Biored: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23(5):bbac282.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022b. Biogpt: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409.

Yubo Ma, Yixin Cao, Yong Hong, and Aixin Sun. 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*, pages 10572–10601, Singapore. Association for Computational Linguistics.

Dat Quoc Nguyen, Karin Verspoor, Trevor Cohn, Lawrence Cavedon, Ameer Albahem, and 1 others. 2020. Overview of ChEMU 2020: Named entity recognition and event extraction of chemical reactions from patents. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2020*, volume 12260 of *Lecture Notes in Computer Science*. Springer, Cham.

Yingjie Peng, Yuanhe Tian, Yan Song, and Fei Xia. 2021. Crossner: Evaluating cross-domain named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

H. Que, J. Liu, G. Zhang, and 1 others. 2024. D-CPT law: Domain-specific continual pre-training scaling law for large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 90318–90354.

Xiaoyu Wang, Yue Yu, Rui Xia, and Min Zhang. 2022. Crossre: A cross-domain dataset for relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3592–3604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xinyu Wang, Wendi Zhou, Chuanqi Zu, Kaiqiang Song, Jingqing Zhang, Bill Yuchen Lin, Xiang Ren, Arman Cohan, and Wenhu Chen. 2023. Instructuie: Multi-task instruction tuning for unified information extraction. *Computing Research Repository*, arXiv:2304.08085.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico. Association for Computational Linguistics.

Zaifu Zhan, Shuang Zhou, Mingchen Li, and Rui Zhang. 2025. Ramie: Retrieval-augmented multi-task information extraction with large language models on dietary supplements. *Journal of the American Medical Informatics Association*, 32(3):545–556.

Qi Zhang, Zhijia Chen, Huitong Pan, Cornelia Caragea, Longin Jan Latecki, and Eduard Dragut. 2024. Scier: An entity and relation extraction dataset for datasets, methods, and tasks in scientific documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA. Association for Computational Linguistics.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2023. Universalner: Targeted distillation from large language models for open named entity recognition. *Computing Research Repository*, arXiv:2308.03279.

A Detailed LLM-Assisted Annotation Pipeline

Candidate Sourcing and Stratification. Our pipeline begins by strategically sourcing and prioritizing candidate passages from a large, unlabeled corpus. This is achieved through a two-step triage process. First, rule-based filtering removes boilerplate using length bounds, language identification, and regex/keyword exclusion of non-content sections. Second, an LLM assigns 1–5 scores for domain relevance, entity richness (density of technical mentions without typed NER), and discourse quality. Passages are then stratified into buckets with fixed sampling ratios to balance value and diversity: a high-value "core" layer (high relevance & richness), a "diversity" layer (high relevance, mid/low richness), and a "negative" layer (low relevance/quality) to expose the model to counterexamples. We retain per-sample metadata (triggered rules, scores, stratum) for later analysis.

Parallel Multi-Model Labeling and Validity Checks. Once promising candidates are identified, they enter our parallel annotation stage, which is designed to generate multiple, diverse extraction outputs for each passage. We configure three *primary* model families and one *backup* family (reserved for arbitration). To foster a spectrum of extraction strategies, each model randomly samples from a library of decoding profiles with distinct parameter sets (e.g., temperature, top-p) that modulate the output from precise to more exploratory. Concurrently, models select from a collection of prompt variants that, while syntactically diverse, all adhere to a unified instructional contract defining the task, schema, and output format. For each passage, this process yields three independent extractions $\{\texttt{result}_1, \texttt{result}_2, \texttt{result}_3\}$. Each output must then pass two strict validators: a JSON schema check and a span-alignment check, which requires every extracted span to be an exact substring of the source text.

Agreement-Based Fusion and Conflict Arbitration With multiple, diverse extraction outputs generated, the next step is to quantitatively assess their consistency and arbitrate disagreements. We employ a conflict-arbitration protocol governed by a controversy score, which determines whether the outputs exhibit sufficient consensus for automated selection or require a tiered resolution process designed to balance data quality, cost, and throughput.

First, we define an equivalence relation $\approx$ (based on alias mapping, unit normalization, etc.). Given the set of items R_i from each result, the common set is $C = R_1 \cap_{\approx} R_2 \cap_{\approx} R_3$, and the controversy score ϕ is calculated as:

$$\phi = \frac{|C|}{\frac{1}{3}\sum_{i=1}^{3} |R_i|}$$

Based on a predefined threshold τ, samples with $\phi \geq \tau$ are deemed low-controversy. For these, we generate an additional candidate, R_{bag}, via majority voting. The purpose of R_{bag} is to synthesize a potentially superior result by combining correct elements from the initial outputs. A Reward Model then selects the single best extraction from the expanded candidate pool $\{R_1, R_2, R_3, R_{\text{bag}}\}$, increasing the probability of obtaining a high-quality label.

Conversely, if $\phi < \tau$, the sample is flagged as high-controversy. This path serves as a cost-effective strategy to salvage ambiguous cases that would otherwise be discarded by a high threshold or accepted with errors by a low one. We selectively invoke a more capable (and costly) *backup* model to produce a fourth result, R_4. The consensus is then re-evaluated across all four outputs. If the sample now meets the threshold, it proceeds to the Reward Model selection; otherwise, it is finally routed to a disagreement pool for definitive human annotation.

Reward-Model Scoring and Thresholded Decisions Following the fusion and arbitration stage, a domain-adapted Reward Model (RM) performs the final quality assessment. Its purpose is to select the optimal candidate from the available set and, crucially, to determine if that candidate's quality is sufficient for it to be accepted as training data.

The RM scores each candidate label set $Y \in \mathcal{Y}$, where $\mathcal{Y}$ is the pool of candidates from the previous step:

$$\mathcal{Y} = \{R_1, R_2, R_3\} \cup \{R_4^{\text{opt}}\} \cup \{R_{\text{bag}}\}$$

Let $s_Y = \text{RM}(Y \mid x)$ be the score for a given candidate. We define two metrics for our decision logic:

$$s_{\max} = \max_{Y \in \mathcal{Y}} s_Y, \quad s_{\text{mean}} = \operatorname*{mean}_{Y \in \mathcal{Y}} s_Y$$

To ensure our RM accurately reflects domain-specific quality preferences, we do not use an open-source RM directly. Instead, we first perform a diagnostic analysis to identify its inherent biases (e.g., preference for longer outputs or higher entity counts). We then mitigate these biases by fine-tuning the RM on a custom-built preference dataset, which drawn from public IE datasets and expert-verified samples.

While an RM excels at ranking responses for a given input, its raw scores are not calibrated to an absolute scale of quality. We therefore establish three empirical decision thresholds $(S_{\text{low}}, S_{\text{mean}}, S_{\text{high}})$ by running the calibrated RM on a reference set of annotations pre-graded by experts as "good," "acceptable," and "bad." S_{low} is set near the upper boundary of the "bad" distribution, S_{high} near the lower boundary of the "good" distribution, and S_{mean} anchors the central mass of the "acceptable" band for cohort-level checks.

The final decision is made using the following rules:

- **Reject:** If the initial candidates are collectively weak ($s_{\text{mean}} < S_{\text{mean}}$) or if even the best candidate is unacceptable ($s_{\max} < S_{\text{low}}$).
- **Acceptable Confidence:** If $S_{\text{low}} \leq s_{\max} < S_{\text{high}}$. The best candidate ($\arg\max$) is kept for auxiliary uses (e.g., populating a retrieval KB) but is excluded from the core SFT dataset.
- **High Quality:** If $s_{\max} \geq S_{\text{high}}$. The best candidate is accepted as high-quality training data.

This threshold-based gating mechanism provides a scalable and systematic method for quality control, ensuring that only high-confidence annotations contribute to the supervised fine-tuning process.

Human annotation, audit, and feedback loop. Human work focuses on two roles: (i) labeling items from the disagreement pool (high controversy or low quality in the previous two steps); and (ii) spot-checking a sample of RM-decided outputs. Human outcomes feed back into threshold calibration, RM fine-tuning, prompt variants, alias/unit maps, and fusion heuristics, forming a closed loop. The final products of this stage are a *high-quality training set* and a *domain knowledge base* for retrieval, both continuously improved by the feedback cycle.