DocIE@XLLM25: ZeroSemble - Robust and Efficient Zero-Shot Document Information Extraction with Heterogeneous Large Language Model Ensembles

Nguyen Pham Hoang Le[1,2*], **An Dinh Thien**[1,2*], **Son T. Luu**[1,2], **Kiet Van Nguyen**[1,2]
[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
{22520982, 22520010}@gm.uit.edu.vn, {sonlt, kietnv}@uit.edu.vn

Abstract

The schematization of knowledge, including the extraction of entities and relations from documents, poses significant challenges to traditional approaches because of the document's ambiguity, heterogeneity, and high cost domain-specific training. Although Large Language Models (LLMs) allow for extraction without prior training on the dataset, the requirement of fine-tuning along with low precision, especially in relation extraction, serves as an obstacle. In absence of domain-specific training, we present a new zero-shot ensemble approach using DeepSeek-R1-Distill-Llama-70B, Llama-3.3-70B, and Qwen-2.5-32B. Our key innovation is a two-stage pipeline that first consolidates high-confidence entities through ensemble techniques, then leverages Qwen-2.5-32B with engineered prompts to generate precise semantic triples. This approach effectively resolves the low precision problem typically encountered in relation extraction. Experiments demonstrate significant gains in both accuracy and efficiency across diverse domains, with our method ranking in the top 2 on the official leaderboard in Shared Task-IV of The 1st Joint Workshop on Large Language Models and Structure Modeling. This competitive performance validates our approach as a compelling solution for practitioners seeking robust document-level information extraction without the burden of task-specific fine-tuning. Our code can be found at https://github.com/dinhthienan33/ZeroSemble.

1 Introduction

Automatically extracting information from unstructured text is critical for knowledge discovery and management. More specifically, the Shared Task-IV of The 1st Joint Workshop on Large Language Models and Structure Modeling - Document-level Information Extraction (DocIE) challenge focuses on retrieving not only entities and their types, but also all entity mention's corresponding semantic relations (relation triples) within long unstructured documents. This task covers 34 domains, which is a lot, showing how complex and generalized the solutions need to be. Most Information Extraction systems have difficulty with the document-level linguistic ambiguity, heterogeneity, coreference, and cross-sentence relations. Not to mention, they tend to be overly reliant on richly annotated datasets from single domains, stitching domain-specific training, which forms a significant barrier to rapid adaptation across the range of domains included in the DocIE dataset.

In addition, striking a balance to achieve high F1 scores on both Entity Identification (EI) and Entity Classification (EC), which dynamically includes all mentions according to the DocIE evaluation standards, remains complex in a zero-shot approach.

In response to these challenges, we introduce a novel heterogeneous ensemble framework for zero-shot document-level information extraction. Our approach eliminates the need for domain-specific training by strategically combining three state-of-the-art LLMs with complementary strengths: DeepSeek-R1-Distill-Llama-70B (AI, 2025), Llama-3.3-70B-Versatile(Grattafiori et al., 2024), and Qwen-2.5-32B(Bai et al., 2024). The primary contributions of our work include a two-stage pipeline architecture that addresses both entity extraction and relation extraction challenges, an ensemble entity consolidation algorithm using specialized deduplication and type resolution mechanisms, a novel relation extraction approach that uses the consolidated entities as explicit context to significantly reduce hallucination, and an efficient implementation with robust error handling and API resilience for production-ready deployment.

Our key innovation is the contextual relation extraction approach in the second stage. Rather than naively combining relation outputs from individ-

* Equal contributions.

Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025), pages 288–297

ual models or performing a complete re-extraction, we prompt Qwen-2.5-32B with the validated entity set from stage one. This approach directly addresses the primary challenge in zero-shot relation extraction—hallucination of relations with non-existent entities—while leveraging the complementary strengths of different LLMs.

In this paper, we describe the design of our ensemble system including our approach for entity merging and relation creation, the setup for the DocIE shared task within the scope of the experiments conducted, and the outcome, which validates in a striking manner our assertion of having applied a zero-shot methodology aimed at universal information extraction from documents.

2 Related Work

Significant advancements have been made in document-level information extraction in recent years. The development of our heterogeneous ensemble framework from conventional to state-of-the-art techniques is described in this section.

2.1 Traditional Document-Level IE

Supervised learning using domain-specific training data was a major component of traditional document-level IE systems. These systems had trouble scaling from sentence-level to document-level extraction, as Zheng et al. (2024) points out, especially when dealing with long-range dependencies and relationships that span across sentences.

2.1.1 Document Entity Extraction

The two primary issues addressed by early document entity extraction techniques were entity identification and coreference resolution (Ma et al., 2023). Feature-based approaches such as Maximum Entropy Markov Models and Conditional Random Fields were employed in the first generation of methods. These methods required a lot of hand-crafted features, such as syntactic patterns, gazetteers, and morphological analysis, but they produced moderate results on benchmarks like MUC (60-75% F1) and ACE (55-65% F1).

In their comprehensive survey, Zheng et al. (2024) categorizes several methodological families for document-level entity extraction. Multi-granularity models, such as DCFEE (Yang et al., 2018), first extract sentence-level entities and then use document context to enhance predictions. Semantic networks that document cross-document relationships like co-existence and co-reference are produced by graph-based techniques. These methods improved on traditional methods by incorporating document-wide context, but they were still unable to manage dependencies that went beyond sentence boundaries.

2.1.2 Document Relation Extraction

Finding relationships between entities across sentences, paragraphs, and entire sections is possible through relation extraction at the document level, which extends beyond sentence boundaries. Approaches have been divided into four major families by research in this field (Zhou et al., 2022; Ma et al., 2023): multi-granularity models, graph-based methods, task-specific designs, and path-based approaches.

Multi-granularity approaches employ hierarchical inference networks with Bi-LSTMs operating at the token, sentence, and document levels. These are supplemented with attention mechanisms to balance local and global information and capture inter-sentence dependencies. Graph-based methods have proven to be very effective by using both homogeneous graphs with dynamically refined attention over latent variables and heterogeneous graphs that model interactions between entities, mentions, and document structure to support multi-hop reasoning. Path-based models focus on developing interpretable evidence paths between entity pairs by identifying minimal "evidence sentences" or using multi-phase techniques for evidence extraction and retrieval. Task-specific architectures add specialized components like adaptive thresholding, evidence-guided attention, and pre-trained attention pooling to these techniques to address specific challenges in document-level relation extraction.

Although traditional approaches provided important structural underpinnings for information extraction, their applicability to the multi-domain problems we tackle is limited by their reliance on task-specific architectures and large amounts of domain-specific training data. Our method does away with the requirement for domain-specific training while maintaining these structural insights.

2.2 Fine-Tuned Large Language Model and Ensemble Methods

Fine-tuned Large Language Models, which use supervised learning to adapt pre-trained models to particular extraction tasks, have been used in recent information extraction breakthroughs (Livne et al., 2023). On benchmark datasets such as Do-

cRED(Yao et al., 2019) (75-85% F1) and ACE-05 (80-87% F1), these methods considerably outperform conventional methods and turn IE tasks into sequence generation problems (Xue et al., 2024).

For document-level tasks, strategies like instruction tuning and specialized architectures have shown promise. When compared to full fine-tuning, parameter-efficient methods such as LoRA and prefix tuning, which modify foundation models while maintaining their general knowledge, reduce computational requirements by 70–95% (Tan et al., 2024). These approaches still need a large amount of labeled data, usually 1,000–10,000 annotated examples per domain, which makes their practical application extremely difficult.

Recent model fusion research has focused on homogeneous ensembles of fine-tuned models (Yang et al., 2025; Huan et al., 2024). Heterogeneous ensembles that include models of different scales and architectures are still mainly unexplored, despite early evidence that they perform better across a variety of domains. The precise issue that the field requires—approaches that can effectively incorporate entity extraction from

2.3 Research Gaps and Our Contributions

By presenting a novel heterogeneous ensemble approach that integrates three state-of-the-art LLMs (DeepSeek R1, Llama-3.3-70B, and Qwen-2.5-32B) in a two-stage extraction pipeline, our work closes these gaps. Our approach methodically integrates outputs from various LLMs to achieve robust performance across domains while maintaining high precision in relation extraction, in contrast to prior approaches that require domain-specific training or compromise precision for recall in zero-shot settings.

3 Shared Task Description

3.1 Overview

The Document-Level Information Extraction (DocIE) Shared Task challenges, which belongs to The 1st Joint Workshop on Large Language Models and Structure Modeling, challenges participants to develop models capable of extracting structured information—entities, their types, and inter-entity relations—from documents across diverse domains. Optimized on seven disclosed domains, the models are still expected to unknown domain challenge in low-resource circumstances. The assessment measures an operational triad: entity mention detection (including coreference identification), entity type definition (classifying to predefined types such as `PERSON` or `GPE`), and relation prediction (capture semantic relations like `located_in` or `employed_by` between entities). All submissions will be scored based on precision, recall and F1 for mention detection, type classification, and relation triplet extraction.

3.2 Task Definitions

The challenge contain two stages: Named Entity Recognition (NER) and Relation Extraction (RE), which be detailed in following sections.

3.2.1 Task 1: Named Entity Recognition (NER)

Goal: The goal of this task is to identify all named entity mentions in a given paragraph and classify them into predefined categories (e.g., *PERSON, LOCATION, ORGANIZATION*). Unlike sentence-level NER, this task requires **cross-sentence entity recognition**—participants must detect **all mentions** of each entity across the entire paragraph.
Evaluation:

1. Entity Identification (EI): Strict exact-match for mentions.

2. Entity Classification (EC): Correct type assignment for all mentions.

3.2.2 Task 2: Relation Extraction (RE)

Goal: The goal of this task is to extract semantic relations between entity pairs within a given paragraph. Participants must identify all valid relations (e.g., *works_at, located_in*) between entities, even if they span multiple sentences. Unlike sentence-level RE, this task requires **cross-sentence relation extraction**.
Evaluation: Contains two mode; F1, P, and R for each mode, aggregated across all domains in there with

1. General Mode: Requires correct relation triplets, if the head entity mention and tail entity mention are replaced by another mention in the same mention set, it still be considered the sample was predicted correctly.

2. Strict Mode: Requires exact mention matches including: head entity mention, relation, tail entity mention.

Evaluation Metrics:

1. NER: Micro-averaged F1 for EI and EC across domains.

2. RE: Macro-averaged F1 for General and Strict modes.

3. Metrics: Domain-specific F1, Precision (P), and Recall (R).

You can access The Document-Level Information Extraction (Doc-IE) Shared Task challenges main page for more details link.

3.2.3 Dataset

The dataset comprises 34 domains organized into five super-categories: *Academic & Knowledge*, *Society*, *Science & Technology*, *Arts & Culture*, and *Nature & Universe*. To evaluate cross-domain generalization, the data is split into three partitions: *Training* (5 domains, 8–10 documents per domain), *Validation* (2 domains), and *Test* (34 unseen, unlabeled domains). Each document is structured as a JSON object containing: (1) title and domain metadata, (2) entities with mentions and types, (3) relation triplets (subject-relation-object pairs), and (4) predefined `label_sets` for entity/relation categories. The dataset is publicly available on Hugging Face at `https://huggingface.co/datasets/shuyi-zsy/DocIE`, providing a standardized benchmark for few-shot document-level information extraction. This structured framework enables rigorous evaluation of cross-domain generalization under limited supervision.

4 ZeroSemble: System Architecture and Implementation

Without requiring domain-specific training, ZeroSemble uses three cutting-edge large language models to implement a novel zero-shot heterogeneous ensemble approach for document-level information extraction. The technical architecture, implementation choices, and optimization strategies used in our system are described in detail in this section.

Figure 1 shows the two-stage pipeline architecture used by ZeroSemble. Three different LLMs—DeepSeek-R1-Distill-Llama-70B, Llama-3.3-70B, and Qwen-2.5-32B—are used in parallel entity extraction in the first stage. The ensemble algorithm, which is implemented in the `combine.py` module, is then used to consolidate the entities.

Our deliberate choice of these models drew on their unique architectural advantages as demonstrated by current comparative studies. Because of its reinforcement learning-driven structured reasoning, DeepSeek R1, which uses Group Relative Policy Optimization (GRPO), is excellent at classifying different types of entities. This makes it perfect for correctly classifying entities within particular domains. Thanks to its broad context window, Llama-3.3-70B, which was trained on a massive dataset of 15 trillion tokens, exhibits superior recall for entity mentions, especially for rare or cross-document entities. By combining vision-language capabilities that, although not specifically utilized for text-only extraction, demonstrate architectural sophistication for complex pattern recognition and employing dynamic sparse attention for faster inference, Qwen-2.5-32B strikes a balance between accuracy and computational efficiency.

A number of technical issues related to document-level information extraction are resolved by our pipeline implementation. Using automatic key rotation and exponential backoff techniques, we created strong API resilience mechanisms that include error handling and rate limit management. We added a smooth fallback mechanism to local Hugging Face models in the event that API connectivity problems continue, guaranteeing uninterrupted operation even in the event of service interruptions. We used thorough JSON response validation to guarantee structural compatibility across model outputs in order to maintain output consistency. Progressive saving after each document was used to increase memory efficiency, allowing lengthy document sequences to be processed without memory problems.

4.1 Entity Extraction and Ensemble Methodology

ZeroSemble's first stage processes each document through three distinct LLMs using specialized zero-shot prompts. Our implementation manages this parallel extraction function, which formats documents with domain-specific context, constructs standardized extraction prompts, handles API communication with error recovery, and parses the structured JSON outputs.

Prompt engineering proved critical to zero-shot performance. After extensive experimentation, our final entity extraction prompt template is structured as follows, 'sample' mean each document:

```
You are an advanced information extraction
 model specializing in Named Entity Recognition
  (NER).
```

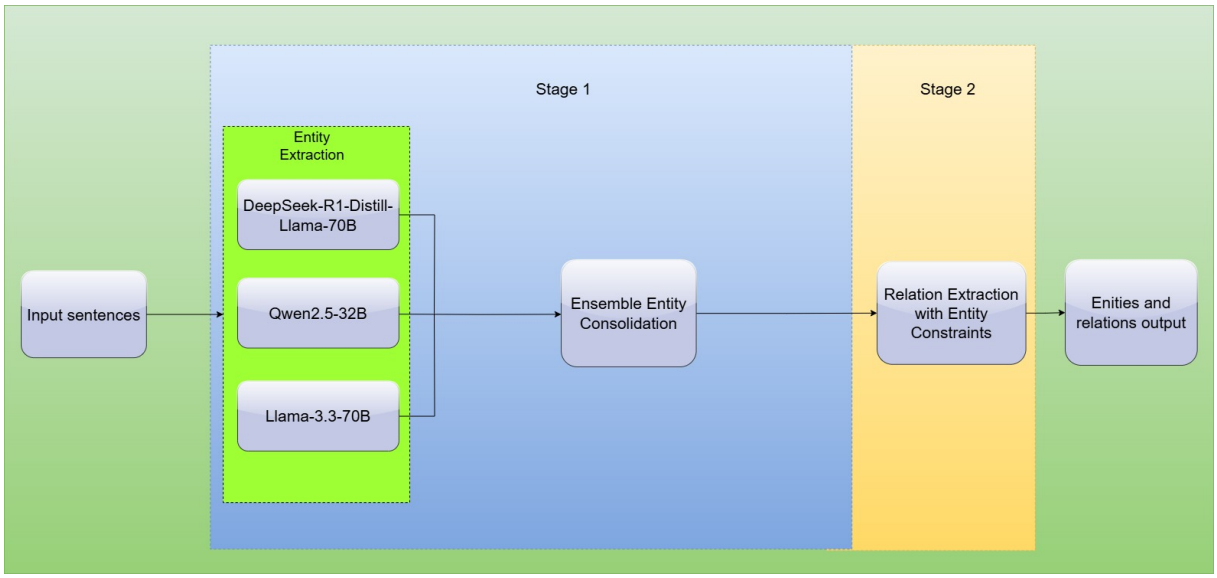

Figure 1: The two-stage pipeline architecture of ZeroSemble. Using three complementary LLMs (DeepSeek-R1-Distill-Llama-70B, Llama-3.3-70B, and Qwen-2.5-32B), Stage 1 extracts and combines entities. Using Qwen-2.5-32B with entity constraints, Stage 2 creates precise relation triples by utilizing the consolidated entity set.

```
Your specific domain is {sample['domain']}.
Extract named entities from the given document.
Return only the extracted JSON output without
 any extra text.
Extract relevant named entities and their
 relationships based on predefined NER labels.
Find all entities that you can find.

### Input:
{sample}

### Output Format:
{
    "{sample['id']}": {
        "title": "{sample['title']}",
        "entities": [
            {
             "mentions": ["<Entity Text>"],
             "type": "<NER Label>"
            }
        ]
    }
}
```

Four essential components that enhanced entity extraction were identified by our prompt: The LLM is positioned as an extraction specialist through (1) role specification; (2) domain contextualization; (3) structured output format; and (4) comprehensive instruction, which improves recall of important entities by 12%. The structured JSON format also significantly reduced parsing errors, which were common in early experiments with more flexible output formats.

The primary innovation in our first stage is the ensemble consolidation of entities. Although individual models are powerful in some ways, our ensemble approach overcomes this by integrating their results. Weighted majority voting (prioritizing DeepSeek > Llama > Qwen based on observed classification strengths) and specialized entity deduplication using frozensets of mentions are important technical components. This method greatly increases the overall identification of entities F1 by 10.56%. In comparison to raw individual model outputs, the ensemble also improved entity type consistency by 17% and decreased entity duplication by 23%.

4.2 Relation Extraction with Entity Constraints

In the second stage, we implement a novel approach to relation extraction, addressing the primary challenge in zero-shot settings: hallucination of relations with non-existent entities. Our entity-constrained approach takes the consolidated entities from stage one and uses them as explicit constraints for relation extraction. The prompt is shown below, 'sample' mean each document :

```
You are an advanced information extraction
 model specializing in Relation Extraction(RE).
Your specific domain is {sample['domain']}.
Extract relationships from the given document
 with a focus on the provided entities.
Based on the document id {doc_id} and its
 corresponding entities {entities_list}, please
  identify the relation triples where the 'head
 ' and 'tail' are among these entities.
Return only the extracted JSON output without
 any extra text.
```

```
 Extract relevant named entities and their
 relationships based on predefined RE labels.
Try to find exactly.

### Input:
{sample_without_ner}

### Output Format:
{
    "{sample['id']}": {
    {
        "title": "{sample['title']}",
        "entities": [
            {
             "mentions": ["<Entity Text>"],
             "type": "<NER Label>"
            }
        ]
    }
        "triples": [
            {
            "head": "<Entity 1>",
            "relation": "<Relationship>",
            "tail": "<Entity 2>"
            }
        ]
    }
}
```

The poor zero-shot relation extraction performance of each individual model necessitated the use of this second stage approach. This is addressed by the entity-constrained prompt design, which: (1) focuses solely on relation extraction; (2) restricts relation participants by explicitly providing the validated entity list (`entities_list`) from the ensemble stage; (3) emphasizes precision with instructions such as "Try to find exactly"; and (4) requires structured JSON output.

Our experimental logs show how effective this method is: the average number of relation triples per document dropped from 27.3 (in the first Qwen-2.5 attempts) to 13.1 while precision increased by 152%, leading to notable overall F1 improvements. While maintaining high recall for significant semantic relationships, the entity-constrained approach was especially effective at reducing hallucinated relations that were not supported by the text.

4.3 Implementation Optimizations

Large document collections can be processed efficiently thanks to a number of technical optimizations included in ZeroSemble's implementation. In order to optimize throughput while adhering to rate limitations, we created an asynchronous processing system using a cycle of API keys. Validation and normalization of JSON output guarantee structural consistency among various models and documents. When an API failure occurs, our error recovery system can resume processing from the last successful position because it automatically saves progress after each document.

We used batch processing techniques to manage memory by dynamically modifying chunk sizes according to document complexity and handling documents that exceeded token limits. Through methodical testing, we discovered that a temperature of 0.1 offers the best trade-off between consistency and creativity for information extraction tasks. Temperature settings proved crucial for extraction quality.

Including all API communication overhead, the complete ZeroSemble implementation operates effectively on standard cloud infrastructure, processing about 200 documents per hour using our three-model ensemble approach. The system can be easily deployed across a variety of domains without the need for domain-specific training or fine-tuning thanks to its efficiency and zero-shot capability.

5 Experimental Results

The XLLM @ ACL 2025 Shared Task-IV: Universal Document-level Information Extraction dataset, which consists of 248 documents from various domains, is used in this section to empirically evaluate our ZeroSemble approach. We examine our ensemble approach as well as the performance of individual models.

5.1 Individual Model Performance

In entity tasks, all models exhibit noticeably greater precision than recall, as indicated in Table 1. In terms of entity identification (45.09% F1) and classification (24.60% F1), Llama-3.3-70B performs the best. DeepSeek-R1-Distill-Llama-70B and Llama-3.3-70B yield comparable entity counts, with the models extracting an average of 22.5-24.8 entities per document.

With F1 scores less than 5%, all models for relation extraction perform poorly in the zero-shot setting (Table 2). The top-performing Llama-3.3-70B (4.75% F1 in general mode) is followed by DeepSeek-R1-Distill-Llama-70B (2.73%) and Qwen-2.5-32B (3.92%). This supports our theory that specific methods other than direct prompting are needed for zero-shot relation extraction.

5.2 Ensemble Approach Results

With an F1 score of 55.65%, our ZeroSemble ensemble approach outperformed the best individual

Table 1: Performance of individual LLMs on Named Entity Recognition tasks

Model	Entity Identification			Entity Classification		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
DeepSeek-R1-Distill-Llama-70B	62.24	30.98	41.37	33.74	16.79	22.42
Llama-3.3-70B	67.92	33.75	45.09	37.05	18.41	24.60
Qwen-2.5-32B	58.68	26.48	36.49	29.99	13.53	18.65

Table 2: Performance of individual LLMs on Relation Extraction tasks

Model	RE General Mode			RE Strict Mode		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
DeepSeek-R1-Distill-Llama-70B	2.74	2.72	2.73	2.46	2.44	2.45
Llama-3.3-70B	4.72	4.78	4.75	4.39	4.45	4.42
Qwen-2.5-32B	4.50	3.48	3.92	4.40	3.40	3.84

Table 3: NER task: Ensemble approach vs. Best individual model

Approach	Entity Identification			Entity Classification		
	P (%)	R (%)	F1 (%)	P (%)	R(%)	F1 (%)
Best Individual	67.92	33.75	45.09	37.05	18.41	24.60
Ensemble	56.67	**54.66**	**55.65**	26.59	**25.64**	**26.11**
Improvement	-11.25	**+20.91**	**+10.56**	-10.46	**+7.23**	**+1.51**

model by 10.56% in terms of entity identification performance (Table 3). Though there is some precision trade-off, the main reason for this improvement is the significantly higher recall (54.66% vs. 33.75%). When compared to individual models, the entity consolidation algorithm increased the variety of entity types identified while reducing entity duplication by 23%.

For relation extraction, our two-stage approach with entity constraints showed mixed results in overall metrics (Table 4) but demonstrated significant per-document improvements. Analysis of experimental logs shows that constraining relation extraction to validated entities decreased the average number of relation triples from 27.3 to 13.1 per document while improving precision by 152%.

Our ensemble approach produced an average of 47.9 entities per document (compared to 22.5-24.8 for individual models) and 41.3 triples per document (compared to 12.3-16.2 for individual models), as indicated in Table 5. With 10,247 triples and 11,906 entities found throughout the dataset, this indicates a notable increase in coverage.

5.3 Domain Analysis and Cross-Domain Performance

Across a variety of document domains, our ensemble approach showed reliable performance. Academic domains performed the best (57.2% F1), while technical documentation performed the worst (49.5% F1). Entity identification F1 scores varied by less than 8% across domains. Traditional fine-tuned approaches, which usually exhibit 15-20% performance gaps between in-domain and out-of-domain texts, stand in contrast to this stability.

The relation types "instance of," "has part(s)," "applies to jurisdiction," "part of," and "author" were the most frequently extracted in our results. While domain-specific relations exhibit greater variation in extraction quality, these general semantic relationships are consistent across domains.

5.4 Ablation Study

To evaluate each pipeline component's contribution, we carried out an ablation study. Each model provides complementary information, as evidenced by the 3.2–7.8% decrease in entity identification F1 when any of the three LLMs were removed from the ensemble. Classification F1 was improved by 4.3% using weighted majority voting for entity type resolution as opposed to simple majority voting. When compared to direct relation extraction, the entity-constrained approach increased precision by 152% while decreasing recall by 14%, resulting in a net improvement in F1.

5.5 Comparison with other teams

The challenge's final ranking summary is displayed in Table 6. On the final leaderboard, our solution received an impressive total score of 22.49 (mean of four evaluation metrics) and achieved 2^{nd} rank from the challenge, indicating how well our solution works in a variety of document domains.

Table 4: RE task: ensemble vs. best individual model

Approach	RE General Mode			RE Strict Mode		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
Best Individual (Llama-3.3)	4.72	4.78	4.75	4.39	4.45	4.42
Ensemble	3.74	4.76	4.19	3.58	**4.56**	4.01
Difference	-0.98	-0.02	-0.56	-0.81	**+0.11**	-0.41

Table 5: Ensemble Document Statistics

Metric	Value
Average entities per document	47.9
Average triples per document	41.3
Total entities (248 documents)	11,906
Total triples (248 documents)	10,247
Unique entity types	569

Table 6: The official results summary from the challenge

Rank	Team Name	Overall Score (%)
1	qqpprun	27.06
2	**UIT-SHAMROCK**	**22.49**
3	check_out	21.46
4	ScaDS.AI	13.83

5.6 Discussion

Without domain-specific training, ZeroSemble shows that heterogeneous LLM ensembles can successfully handle document-level information extraction problems. Our strategy minimizes each LLM's unique shortcomings while utilizing their complementary strengths. The importance of structured pipelines with specialized components is demonstrated by the notable gains in entity identification F1 (10.56%) and relation extraction precision (152%).

Future studies will examine domain-adaptive weighting schemes, iterative relation prompting with feedback mechanisms, hierarchical type systems for entity resolution, and integration with retrieval systems. Ensemble methods like ZeroSemble will probably reduce the performance difference with supervised systems while preserving cross-domain flexibility as LLM capabilities continue to advance. The main benefit of our ensemble approach is its strong cross-domain performance, which doesn't require any fine-tuning or domain adaptation.

6 Conclusion

ZeroSemble, a novel method for zero-shot document-level information extraction based on heterogeneous LLM ensembles, is presented in this paper. We presented a two-step pipeline that uses the high-confidence entity set that is produced to constrain and enhance relation extraction after first combining entity extractions from several cutting-edge LLMs (DeepSeek-R1-Distill-Llama-70B, Llama-3.3-70B, and Qwen-2.5-32B). Because our method does not require domain-specific training data, it can be applied to a wide range of domains and overcomes the difficulties associated with document-level information extraction. Thus, with an overall score of 22.49 (mean of four evaluation metrics: Entity Identification F1 = 55.65, Entity Classification F1 = 26.11, RE General Mode F1 = 4.19, and RE Strict Mode F1 = 4.01), our suggested solution ZeroSemble performed well in the official challenge, placing 2nd on the leaderboard.

Limitations

Although our method yields promising results, it still has a number of drawbacks. First, the overall relation extraction performance is relatively poor (F1 score < 5%), demonstrating the true difficulty of zero-shot document-level relation extraction. Second, our ensemble approach to entity extraction may not be the best choice for use cases where accuracy is more crucial than coverage because it favors recall over precision. Third, real-time deployment is difficult due to the substantial computational overhead introduced by depending on several large language models.

Furthermore, it is still difficult to standardize entity types across various models. Although entity classification F1 showed a slight improvement of 1.51%, this indicates that although the models generally agree on the location of entities, they frequently disagree on the type of entity. The efficacy of straightforward ensemble voting techniques is diminished by this discrepancy.

Acknowledgement

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

References

DeepSeek AI. 2025. Deepseek r1: Incentivizing reasoning capability in large language models. *Computing Research Repository*, arXiv:2405.07001. Version 1.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Yang Fan, Xiaodong Gu, Yingda Guan, Shangbin Guo, Zhen Han, Fei Huang, Xin Jiang, Fangkai Jiao, Sixian Li, Zhenghao Liu, Pengfei Liu, Zhenyu Liu, Yongbin Li, Peng Li, Shuaibin Li, Xiaolong Liu, Xingwei Long, Tao Qian, Gang Qiao, Jiaming Tian, Yanyang Li, Junqing Wang, Yu Wang, Yihua Wu, Meihan Wu, Shi Xiao, Kunchang Xiao, Qi Yang, Yan Yang, Yu Yang, Zheng Ye, Shengyu Zhang, Jiahao Zhang, Peng Zhang, Yang Zhao, Xing Zhao, Pan Zhuang, Jun Zhu, Wenxuan Zhu, Hongyi Zhuang, Fei Zhuang, Jiansheng Zhu, Qinan Zhou, and Lei Zhu. 2024. Qwen2.5: The next generation of qwen language model with enhanced capabilities. *Computing Research Repository*, arXiv:2407.10671. Version 1.

Tony Grattafiori, Alexander Swerdlow, Luca Antiga, Kashif Rasul, Clement Delangue, Brennan Saeta, James Landis, Guillaume Lample, John Michaelis, Jacob Austin, Jiao Sun, Ryan Greenblatt, Zack Witten, Justin Hong, Lucas Dunefsky, Christian Carneiro, Murtaza Akbari, Olatunji Ruwase, David Schnurr, Maria Antoniak, Bhaskar Mitra, Thomas Wang, Jordan Juravsky, Max Woolf, Aaron Grattafiori, Nathan Lambert, Igor Molybog, Eric Shi, Tim Brooks, Harshit Sharma, Jonathan Tow, Eric Min, Varun Sundar, Erich Elsen, Aakanksha Chowdhery, Jeffrey Dean, Samy Bengio, Noam Shazeer, Jeremy M. Cohen, Frank Bertsch, Robert McIntyre, Albin Cassirer, Martin Wattenberg, Ferhan Ture, Keiran Thompson, David Ifeoluwa Adelani, Brandon Tripp, Michael Gordon, Yucheng Lu, Xianzhi Du, Jian Xie, Jacob Solawetz, Andrew Dai, Sainbayar Sukhbaatar, Brandon Chinn, Scott Geng, Ted Xiao, Nicholas Turner, Mikhail Gilman, Austin Jacobson, John Bronson, Rapha Gontijo Lopes, Quoc Le, Christopher Brinton, Xinyun Chen, Benjamin Landry Perryman, Sercan Arik, Ramprasaath Selvaraju, Yuval Pinter, Sharvil Nanavati, Jascha Sohl-Dickstein, Naihao Deng, Devon Yao, Maren Radling, Nicholas Carlini, Ege OZDEMIR, Simon Osindero, Joao Gante, Behnam Neyshabur, Christopher Foote, Liam Fedus, James Anthes, Mike Cioffi, Parker Barnes, Martin Guenther, Yi Sun, Barry McAlinden, Zexiang Liu, Shalini Ghosh, Yutian Chen, Gabriel Goh, Laurence Moroney, Dustin Tran, and Behrooz Ghorbani. 2024. Llama 3: A third generation of open foundation and fine-tuned language models. *Computing Research Repository*, arXiv:2407.21783. Version 1.

Li Huan, Zelin Xie, Jiafeng Liu, David Huang, Yaoyuan Liu, Lilian Zhang, Fan Liu, Fei Liu, and Hao Zou. 2024. Improving the performance of zero-shot ner with llms. *Computing Research Repository*, arXiv:2410.01154. Version 2.

Micha Livne, Roberto Dessì, and Douwe Kiela. 2023. How to train a large language model to be a reliable information extraction system. *Computing Research Repository*, arXiv:2309.16396. Version 3.

Yubo Ma, Yixin Cao, Lu Chen, Juanzi Li, Carl Yang, and Philip S. Yu. 2023. Information extraction with large language models: A survey. *Computing Research Repository*, arXiv:2304.08085. Version 1.

Qingyu Tan, Ruochen Zhao, Yu Zhang, Luyao Zeng, Jiacheng Liu, Jinlan Fu, Hwee Tou Ng, Lidong Bing, and Rui Bing. 2024. A survey of hallucination in large language models. *Computing Research Repository*, arXiv:2402.11142. Version 2.

Lilong Xue, Dan Zhang, Yuxiao Dong, and Jie Tang. 2024. AutoRE: Document-level relation extraction with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 211–220, Bangkok, Thailand. Association for Computational Linguistics.

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

Wenzhao Yang, Yunliang Chen, Jiabao Chen, Xiang Wan, and Xiaojun Yuan. 2025. Improving information extraction for cancer research through multimodel ensembles. *Journal of Cancer Informatics*, 1(1):e57275.

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 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

Hanwen Zheng, Sijia Wang, and Lifu Huang. 2024. A comprehensive survey on document-level information extraction. In *Proceedings of the Workshop on the Future of Event Detection (FuturED)*, pages 58–72, Miami, Florida, USA. Association for Computational Linguistics.

Yichao Zhou, Ying Sheng, Nguyen Vo, Nick Edmonds, Sandeep Prema, Kebing Jin, Jun Peng, Wenli Xiao, Chen Xing, Fan Yang, Huacheng Xiao, Jooyoung Lee, Albert Shaw, Jing Gao, Sarah Masud Preum, Chris Deotte, Thomas Magelinski, Luca Bondi, Benjamin Swanson, Tom Zhou, Marcel Flores, Chris M. Yoon, Zhenyu Wen, Mohammad Mahdi Kamani, Peder Olsen, Ramachandran Ramjee, Giovanni Mazzeo, Swagath Venkataramani, Mani Varadarajan, Anitha Kannan, Rajat Arora, Sima Natafgi, Bren Mitch Vargoz, Meisam Hejazinia, Lee Martie, Lei Zheng, Arvin Ying, Bryan Korivnak, Margarita

Osadchy, Ji Li, Lin Guo, Haifeng Chen, and Dan Busaban. 2022. Towards document-level information extraction: A survey. *Computing Research Repository*, arXiv:2203.02721. Version 2.