Understanding the Impact of Coreference Resolution and Structured Prompts in LINK-KG

Dipak Meher
Email: dmeher@gmu.edu
CS 782 Advanced-ML Project Report
GitHub: https://github.com/dipakmeher/cs782AdvancedML

I. Introduction

U.S.-bound migration has recently undergone significant transformation. The rise in irregular border crossings and asylum applications has amplified the role of human smuggling networks. These networks operate across transnational corridors and are increasingly sophisticated, often forming alliances with organized criminal groups. Although these operations are illegal, they function through intricate webs of interconnected actors and entities. This raises a compelling question: if such networks are inherently interconnected, why not represent them as Knowledge Graphs (KGs) and leverage them for downstream tasks such as identifying key actors, detecting operational patterns, and analyzing temporal evolution?

Recent work such as LINK-KG [1], a modular and Large Language Model (LLM)-driven framework, has shown effectiveness in constructing accurate and interpretable knowledge graphs from legal case documents related to human smuggling. LINK-KG employs a cache-based coreference resolution pipeline to disambiguate varying entity mentions (e.g., "Defendant John Smith," "Smith," "Defendant"), resolve role shifts (e.g., a smuggler later referred to as a driver), ambiguous aliases (e.g., "the agent" referring to different individuals), and plural references (e.g., "the agents" denoting multiple people). It then uses structured prompting strategies for entity and relationship extraction, including in-prompt noise filtering, sequential type-wise extraction to reduce attention drift, and explicit type definitions to prevent overgeneralization. This framework has demonstrated significant improvements in reducing node duplication and legal noise.

II. Problem Definition

This project aims to rigorously analyze the contribution of two fundamental components within the LINK-KG framework: (1) the cache-based coreference resolution mechanism, and (2) the modified structured prompting strategy for entity and relationship extraction.

The first objective is to conduct a controlled ablation study that isolates each component to quantify its individual impact on knowledge graph construction. In particular, we evaluate how the removal of each module affects graph quality, with emphasis on duplication rates, noise reduction, and overall structural coherence.

The second objective is to evaluate how effectively structured prompts guide an LLM to extract entities and relationships with high precision and recall. By evaluating extraction performance without using any post-processing or correction modules, we assess how well structured prompting alone supports reliable knowledge graph generation and where its limitations appear.

III. Methodology

A. Ablation Study: Coreference Resolution and Structured Prompting in LINK-KG

Figure 1 presents an overview of the four pipeline configurations evaluated in our ablation study. The first configuration corresponds to the full LINK-KG framework, which integrates both the cache-based coreference resolution module and the modified structured prompts for entity and relationship extraction. This serves as the reference model against which all variants are compared.

The second configuration, LinkKG-no-coref, removes the coreference resolution module. In this setting, the legal text is passed directly to the extraction component guided by the structured prompts, followed by knowledge graph construction. This variant enables the isolation of the contribution of coreference resolution to downstream KG quality.

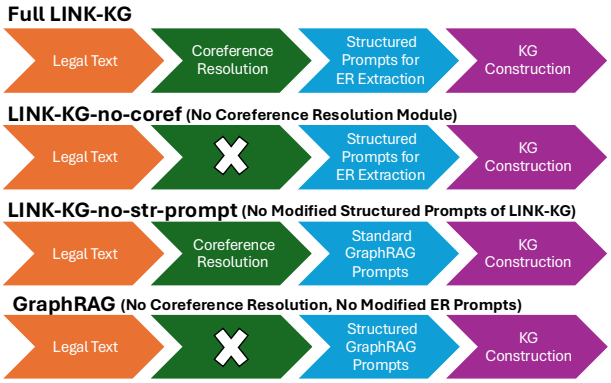


Fig. 1: Ablation setups for LINK-KG.

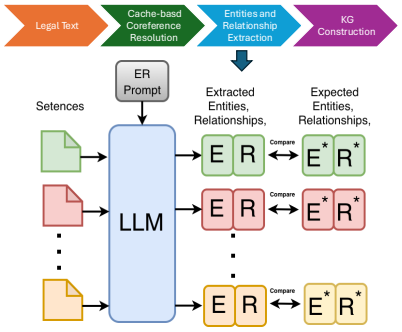


Fig. 2: Entity and relationship extraction workflow.

The third configuration, LinkKG-no-str-prompt, excludes the modified structured prompts while retaining the coreference module. Here, the extraction process relies on the standard GraphRAG prompts [2], a retrieval-augmented generation framework that leverages graph structures to enhance contextual relevance. To maintain comparability across all variants, the KG construction module remains identical to that used in the full LINK-KG framework.

The fourth configuration corresponds to the baseline GraphRAG pipeline, which removes both the coreference resolution module and the modified structured prompts. This setup represents a minimal KG construction process executed without any refinement mechanisms, providing a reference point for evaluating the individual contributions of the additional components in the full LINK-KG pipeline.

Together, these configurations allow us to systematically assess the individual and combined impact of coreference resolution and structured prompting on the quality, coherence, and noise characteristics of the constructed knowledge graphs.

B. Entity and Relationship (ER) Evaluation

This evaluation assesses the effectiveness of the structured prompt used in LINK-KG for guiding entity and relationship extraction. The objective is to evaluate whether the LLM can accurately identify and extract key entities and relationships from legal case text. Given a set of sentences from the legal corpus, the model's predicted entities and relationships are compared against ground-truth annotations to measure extraction accuracy and precision. The findings from this evaluation provide insight into the strengths and limitations of the current ER extraction strategy and inform the design of improved prompting methods for future iterations of the framework.

IV. Experiments

A. Dataset

For the ablation study, we use legal case documents related to human smuggling obtained from the Nexis Uni database. All ablation settings (Figure 1) are evaluated on the same set of 16 legal cases, consisting of both short and long documents, consistent with those used in the original LINK-KG framework [1].

For the NER evaluation, we use a manually annotated dataset that includes: (i) predefined entity types, (ii) corresponding sentences from legal case documents, and (iii) ground-truth entities and relationships for each sentence. We use the same seven entity types as in the LINK-KG framework: *Person*, *Location*, *Routes*, *Organization*, *Means of Transportation*, *Means of Communication*, and *Smuggled Items*. The dataset currently consists of approximately 541 annotated samples. Table I reports the overall dataset statistics.

Category	**Count**
Entities	1536
Relations	1183
Per-Type Distribution	
Person	668
Means of Transportation	262
Means of Communication	15
Routes	64
Location	371
Smuggled Items	117
Organization	25
Total	**1536**

TABLE I: Summary of annotated entities, relations, and per-type entity distribution in the ER evaluation dataset.

Case	Node Duplication												Noisy Nodes											
	GraphRAG			LinkKG-no-coref			LinkKG-no-str-prompt			LinkKG			GraphRAG			LinkKG-no-coref			LinkKG-no-str-prompt			LinkKG		
	Tot	Dup	Rate	Tot	Dup	Rate	Tot	Dup	Rate	Tot	Dup	Rate	Tot	Noisy	Rate	Tot	Noisy	Rate	Tot	Noisy	Rate	Tot	Noisy	Rate
Case 01	94	32	34.04	58	18	31.03	75	20	26.67	46	7	15.22	94	26	27.66	58	4	6.90	75	17	22.67	46	3	6.52
Case 02	86	24	27.91	59	10	16.95	79	17	21.52	52	7	13.46	86	28	32.56	59	3	5.08	79	25	31.65	52	2	3.85
Case 03	60	19	31.67	25	5	20.00	35	6	17.14	22	3	13.64	60	17	28.33	25	7	28.00	35	10	28.57	22	6	27.27
Case 04	75	15	20.00	41	6	14.63	57	10	17.54	34	2	5.88	75	19	25.33	41	7	17.07	57	12	21.05	34	5	14.71
Case 05	49	11	22.45	23	4	17.39	31	4	12.90	21	3	14.29	49	9	18.37	23	4	17.39	31	10	32.26	21	4	19.05
Case 06	68	20	29.41	49	13	26.53	64	8	12.50	32	2	6.25	68	10	14.71	49	4	8.16	64	13	20.31	32	2	6.25
Case 07	55	13	23.64	38	5	13.16	61	16	26.23	36	2	5.56	55	10	18.18	38	3	7.89	61	10	16.39	36	3	8.33
Avg	69.57	19.14	27.02	41.86	8.71	19.96	57.43	11.57	19.21	34.71	3.71	10.61	69.57	17.00	23.59	41.86	4.57	12.93	57.43	13.86	24.70	34.71	3.57	12.28

TABLE II: Comparison of node duplication and legal noise across different extraction pipelines for small legal case documents. *Tot* denotes the total number of extracted entities, *Dup* denotes the number of duplicated entities, and *Noisy* denotes the number of noisy entities. *Rate* indicates the percentage with respect to the total.

B. Evaluation Metrics

For the ablation study, we adopt the same evaluation metrics used in LINK-KG: *node duplication rate* and *noise rate*. The node duplication rate measures the proportion of duplicated entity nodes in the constructed knowledge graph, while the noise rate measures the proportion of irrelevant legal or government-related entities present in the graph.

The node duplication rate is defined as:

$$\text{Duplication Rate} = \frac{\text{Number of Duplicate Nodes}}{\text{Total Number of Nodes}} \tag{1}$$

The noise rate is defined as:

$$\text{Noise Rate} = \frac{\text{Number of Noisy Nodes}}{\text{Total Number of Nodes}} \tag{2}$$

For the NER evaluation, we use Precision, Recall, and F1-score as the primary evaluation metrics. Among these, the F1-score is the main metric, as it provides a balanced measure of precision and recall, and is well-suited for evaluating overall extraction quality. The F1-score is defined as:

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

C. Experimental Setup

For both the ablation study and the ER evaluation, we use the open-source Llama 3.1 70B model, accessed locally through Ollama. For the GraphRAG baseline, we configure a chunk size of 300 tokens and use the same seven entity types defined in LINK-KG: *Person*, *Location*, *Routes*, *Organization*, *Means_of_Transportation*, *Means_of_Communication*, and *Smuggled_Items*.

For the ER evaluation, we use the modified structured prompts in LINK-KG, referred to as *Structured Prompts*, and the prompts used in GraphRAG, referred to as *GraphRAG Standard Prompts*.

V. RESULTS

A. Ablation Study

Tables II and III report the complete ablation results across all 16 legal case documents, consisting of 7 short and 9 long cases. The results are presented for both evaluation metrics: node duplication rate and noise rate.

1) Node Duplication: As shown in Figure 3, removing the coreference resolution module leads to a notable increase in duplicate nodes. **LinkKG-no-coref** exhibits a duplication rate of 24.05%, compared to 14.20% in LINK-KG, reflecting a relative increase of 69.4%. The primary reason is the absence of the coreference resolution module, which consolidates duplicate variants of entity mentions, ensures consistency, and helps remove duplication. This also emphasizes that structured prompts alone are insufficient to reduce duplication in knowledge graphs.

LinkKG-no-str-prompt exhibits a duplication rate of 21.21%, which is a relative increase of 49.4% compared to LINK-KG. While this is lower than LinkKG-no-coref due to the presence of the coreference resolution module, the

Case	Node Duplication												Noisy Nodes											
	GraphRAG			LinkKG-no-coref			LinkKG-no-str-prompt			LinkKG			GraphRAG			LinkKG-no-coref			LinkKG-no-str-prompt			LinkKG		
	Tot	Dup	Rate	Tot	Dup	Rate	Tot	Dup	Rate	Tot	Dup	Rate	Tot	Noisy	Rate	Tot	Noisy	Rate	Tot	Noisy	Rate	Tot	Noisy	Rate
Case 08	83	20	24.10	51	11	21.57	74	10	13.51	40	5	12.50	83	54	65.06	51	14	27.45	74	41	55.41	40	10	25.00
Case 09	131	50	38.17	82	24	29.27	109	21	19.27	80	15	18.75	131	33	25.19	82	16	19.51	109	23	21.10	80	16	20.00
Case 10	104	37	35.58	74	19	25.68	100	20	20.00	50	10	20.00	104	26	25.00	74	10	13.51	100	25	25.00	50	6	12.00
Case 11	149	56	37.58	70	16	22.86	150	45	30.00	64	12	18.75	149	58	38.93	70	15	21.43	150	53	35.33	64	4	6.25
Case 12	171	65	38.01	88	33	37.50	120	41	34.17	45	8	17.78	171	125	73.10	88	38	43.18	120	60	50.00	45	17	37.78
Case 13	183	59	32.24	79	24	30.38	166	32	19.28	63	9	14.29	183	104	56.83	79	18	22.78	166	76	45.78	63	15	23.81
Case 14	158	62	39.24	79	16	20.25	105	33	31.43	62	8	12.90	158	28	17.72	79	12	15.19	105	12	11.43	62	6	9.68
Case 15	99	28	28.28	60	16	26.67	87	14	16.09	46	8	17.39	99	38	38.38	60	21	35.00	87	41	47.13	46	6	13.04
Case 16	214	109	50.93	133	52	39.10	131	34	25.95	76	21	27.63	214	56	26.17	133	24	18.05	131	43	32.82	76	8	10.53
Avg	143.56	54.00	36.01	79.56	23.44	28.14	115.78	27.78	23.30	58.44	10.67	17.78	143.56	58.00	40.71	79.56	18.67	24.01	115.78	41.56	35.90	58.44	9.78	17.57

TABLE III: Comparison of node duplication and legal noise across different extraction pipelines for long documents. *Tot* denotes the total number of extracted entities, *Dup* denotes the number of duplicated entities, and *Noisy* denotes the number of noisy entities. *Rate* indicates the percentage with respect to the total.

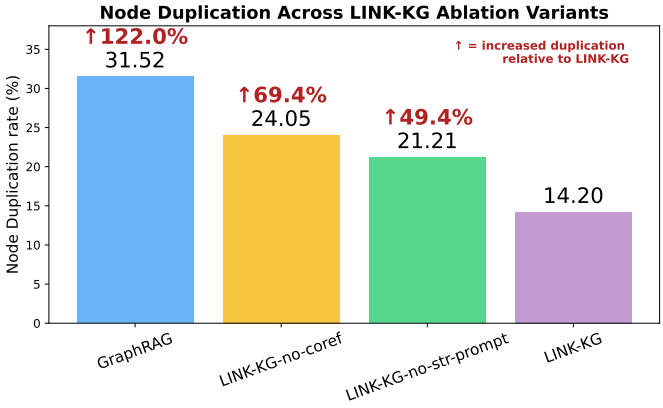


Fig. 3: Node Duplication Across LINK-KG Ablation Variants

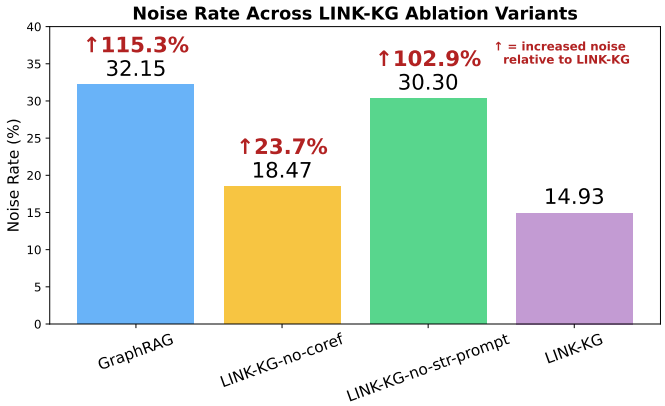


Fig. 4: Noise Rate Across LINK-KG Ablation Variants

increase is still significant. This is mainly because LinkKG-no-str-prompt uses the standard GraphRAG prompts, which are less constrained and lead to higher entity extraction. Figure 5 illustrates this trend, showing that LinkKG-no-str-prompt extracts the highest number of entities compared to both LinkKG-no-coref and LINK-KG. This increase in extraction also introduces a large number of duplicate nodes, which raises the overall duplication count despite the use of coreference resolution.

GraphRAG shows the highest duplication rate at 31.52%, underscoring the effectiveness of both coreference resolution and structured prompting in minimizing redundancy. Since GraphRAG does not employ either coreference resolution or the structured prompts used in LINK-KG, it exhibits a substantial increase in node duplication, exceeding 100% relative to LINK-KG, as shown in Figure 3. This provides two key insights. First, the coreference resolution module plays a critical role in reducing duplication. Second, structured prompts also contribute to duplication reduction, as seen by comparing GraphRAG with LinkKG-no-coref: although both lack coreference resolution, LinkKG-no-coref achieves a lower duplication rate of 24.05% due to the use of structured prompts, compared to 31.52% for GraphRAG. This corresponds to a relative increase of 31.06% in GraphRAG. The primary reason is that higher precision in extraction also reduces the extraction of loose or irrelevant entities and their associated duplications.

The **head-to-head comparison** is shown in Figures 6 and 7. LINK-KG consistently performs better in reducing duplication across both short and long documents. For short documents, compared to LinkKG-no-coref and LinkKG-no-str-prompt, there is an improvement of 46.8% and 44.8%, respectively. For long documents, the corresponding improvements are approximately 36.7% and 23.4%, respectively. GraphRAG shows more than a 50% increase in duplication for both short and long documents. This consistent improvement highlights the impact of coreference resolution in reducing duplication.

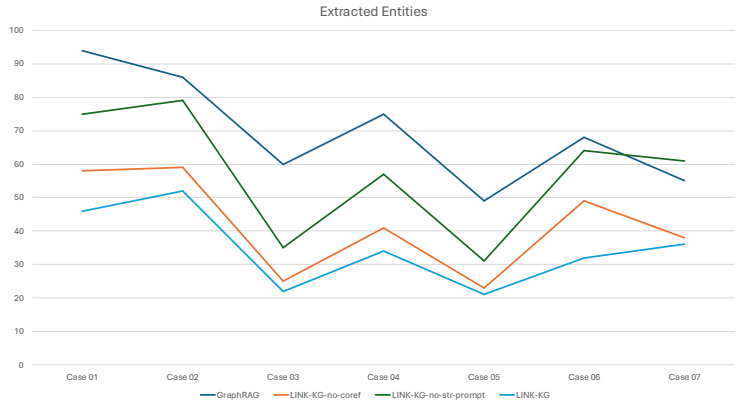


Fig. 5: Total extracted entities in small documents across different variants. LinkKG-no-str-prompt shows a higher number of extracted entities due to the absence of structured prompts used in LINK-KG.

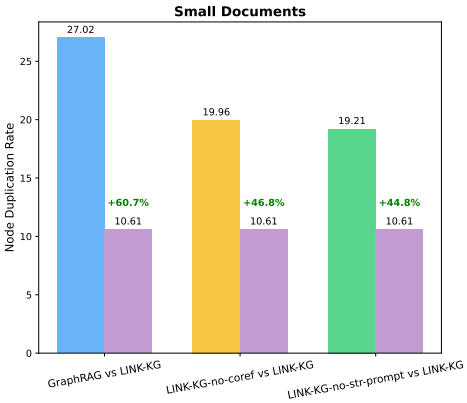


Fig. 6: Node Duplication for Short Documents

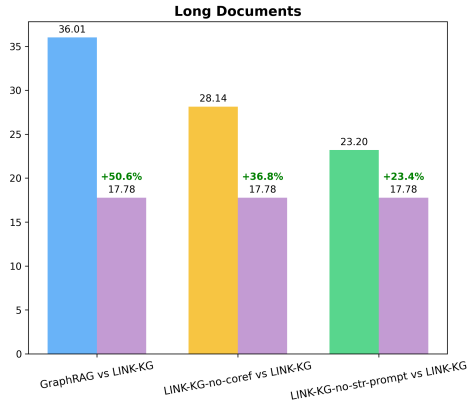


Fig. 7: Node Duplication for Long Documents

2) Noise Reduction: As shown in Figure 4, removing the structured prompts leads to a substantial increase in legal noise. **LinkKG-no-str-prompt** exhibits a noise rate of 30.30%, compared to 14.93% in LINK-KG, reflecting a relative increase of 102.9%. The primary reason is the absence of structured prompts, which perform in-prompt filtering to suppress irrelevant entities and apply sequential extraction to reduce attention drift, enabling more precise entity and relationship extraction. Although this variant still includes the coreference resolution module, the increased noise indicates that coreference alone is not sufficient for noise reduction. Instead, structured prompting plays the dominant role in controlling legal noise.

LinkKG-no-coref exhibits a noise rate of 18.7%, which is a relative increase of 14.93% compared to LINK-KG. This is a relatively smaller increase in noise rate, primarily because both use structured prompts, which reduce legal noise during extraction.

GraphRAG shows a significant increase in noise of around 115.3%, due to the absence of both coreference resolution and structured prompts. This also indicates that the coreference resolution module has some influence in reducing noise, since LinkKG-no-str-prompt, which includes coreference resolution, has a noise rate of 30.30%, which is lower than GraphRAG at 32.15%. This corresponds to a relative reduction of 6%. Although this reduction is small, the primary reason is that loose entity extraction increases the total number of extracted entities, which overshadows the gains obtained from coreference resolution for noise reduction. While the coreference resolution module consolidates legal noise variants, loose extraction also introduces additional irrelevant entities that increase the overall legal noise count. For example, *Western District of Texas Court*, *District Court*, *Court*, and *Western District* are extracted in Case 9 as separate entities due to loose extraction, even though they refer to the same

underlying entity. This directly increases the legal noise count.

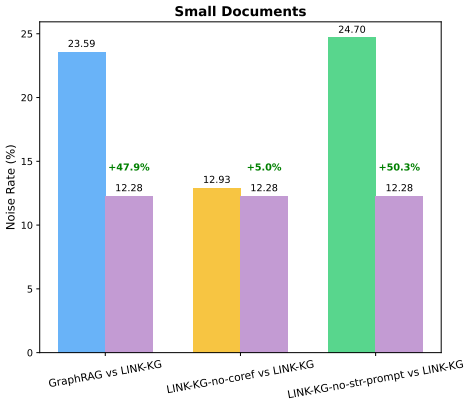


Fig. 8: Noise Rate for Short Documents

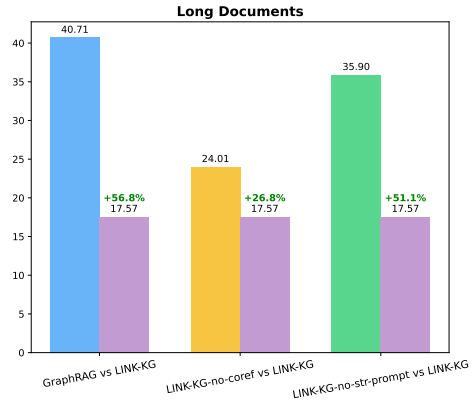


Fig. 9: Noise Rate for Long Documents

The **head-to-head comparison** is shown in Figures 8 and 9. LINK-KG consistently performs better in reducing noise across both short and long documents. For short documents, compared to LinkKG-no-coref and LinkKG-no-str-prompt, there is an improvement of 5% and 50.3%, respectively. For long documents, the corresponding improvements are approximately 26.8% and 51.1%, respectively. GraphRAG shows around a 50% increase in noise for both short and long documents. This consistent trend highlights the impact of structured prompts in reducing legal noise.

B. Qualitative Analysis of Extracted Graphs

Figure 10 presents the generated graphs from **LINK-KG**, **LinkKG-no-coref**, and **LinkKG-no-str-prompt** for Case 6. This section analyzes the structural gains achieved through the incorporation of coreference resolution and structured prompts in the LINK-KG pipeline. The node sizes are scaled based on their degree (i.e., the number of relationships connected to each node).

1) Analysis of Relationship-to-Node Ratio: Table IV presents the node count, relationship count, and relationship-to-node ratio (R/N) for graphs generated by LINK-KG, LinkKG-no-coref, and LinkKG-no-str-prompt. To assess the structural quality of these graphs, we compute the R/N ratio, which reflects how densely the nodes are interconnected.

TABLE IV: Graph Statistics for Case 6

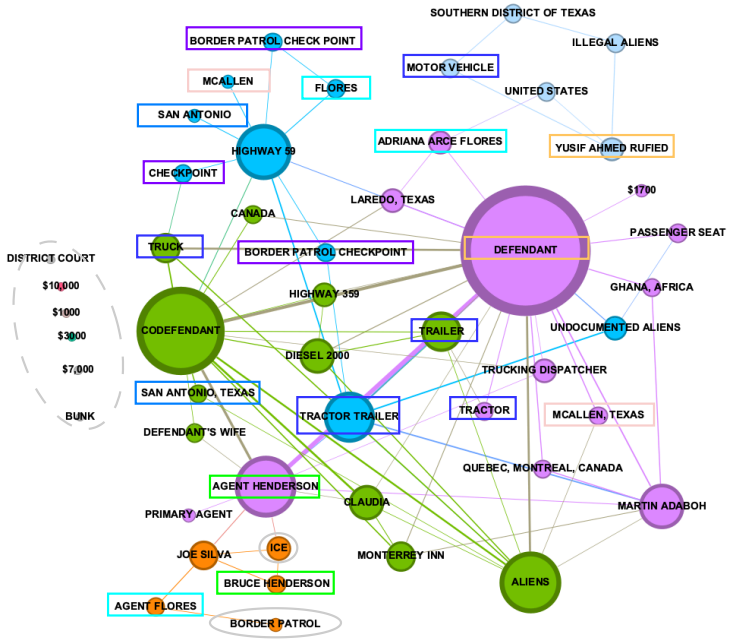
Method	#Nodes	#Relationships	R/N Ratio
LINK-KG	32	71	2.21
LinkKG-no-coref	49	86	1.75
LinkKG-no-str-prompt	64	100	1.56

The graph constructed using LinkKG-no-str-prompt contains the highest number of nodes and edges, with 64 nodes and 100 edges, indicating extensive but less controlled extraction due to the use of GraphRAG standard prompts. LinkKG-no-coref generates a medium-scale graph with 49 nodes and 86 edges, while LINK-KG produces the most compact graph, consisting of only 32 nodes and 71 edges.

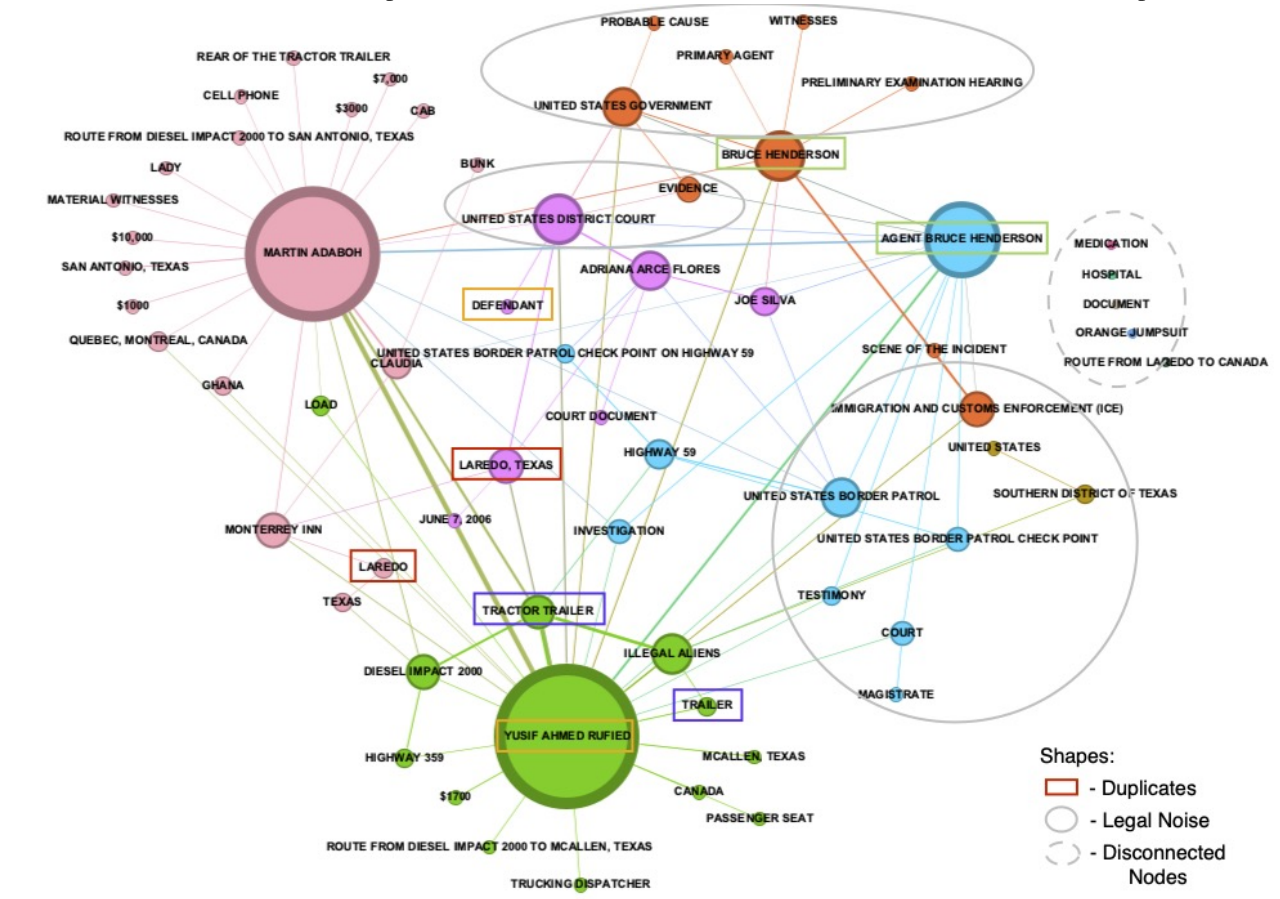
Among all variants, LINK-KG attains the highest relationship-to-node (R/N) ratio of 2.21, reflecting stronger structural coherence. In comparison, LinkKG-no-coref and LinkKG-no-str-prompt achieve lower R/N ratios of 1.75 and 1.56, respectively. Although LinkKG-no-str-prompt extracts a larger number of entities, many of these nodes exhibit weak or isolated connections, indicating higher noise. The increased density and compactness of the LINK-KG graph highlight the combined effectiveness of coreference resolution and structured prompting in generating a more coherent and informative knowledge graph.

2) Per-Case Graph Comparison and Structural Gains: We further examine the extracted graphs through visual inspection to analyze the impact of duplicate and noisy nodes. In Figure 10, duplicate nodes are highlighted using rectangles, while noisy nodes are enclosed within ovals. Rectangles with identical colors indicate different surface forms referring to the same real-world entity, thereby representing duplication.

(a) LINK-KG Graph



(b) LinkKG-no-coref Graph



(c) LinkKG-no-str-prompt Graph

Fig. 10: Graphs of (a) LINK-KG; (b) LinkKG-no-coref; (c) LinkKG-no-str-prompt, shown for Case 06. Duplicate nodes are indicated with solid rectangles, noisy or irrelevant nodes with solid ovals, and disconnected nodes connected nodes with dashed ovals.

a) Analysis of LinkKG-no-coref Graph: Figure 10b presents the graph generated using LinkKG-no-coref. As shown in Table II, this variant produces a total of 49 nodes, which is slightly higher than LINK-KG (32) and substantially lower than LinkKG-no-str-prompt (64). However, it exhibits a noticeably higher duplication rate (26.53%) compared to LINK-KG (20.27%) and LinkKG-no-str-prompt (12.50%). This increase is primarily due to the absence of the coreference resolution module, which fails to unify multiple surface forms of the same entity, causing them to be treated as distinct nodes.

For instance, in the LinkKG-no-coref graph (Figure 10b), the main defendant, "Yusif Ahmed Rufied," appears under multiple surface variations such as "Defendant." Similarly, other person-type entities, including "Andriana Arce Flores" and "Bruce Henderson," are each represented as multiple nodes due to unresolved duplicate mentions. Location-type entities such as "San Antonio, Texas" appear separately as "San Antonio." Likewise, means of transportation such as "Truck" are scattered across multiple forms including "Motor Vehicle," "Tractor Trailer," "Tractor," and "Trailer." These duplications, among many others, introduce structural redundancy and reduce graph clarity, highlighting the importance of applying coreference resolution prior to knowledge graph construction.

With respect to noise, LinkKG-no-coref performs better than LinkKG-no-str-prompt and remains comparable to LINK-KG. The structured prompts still filter many irrelevant entities and support relatively precise extraction. However, several entities remain disconnected from the main graph, as shown in Figure 10b. This occurs mainly because some nodes lack extracted relationships even when those relations are present in the text. This behavior arises due to attention drift, where the model processes many entities within dense sentences and overlooks certain links, leaving those nodes isolated. Since LinkKG-no-coref does not apply coreference resolution, the presence of multiple surface variants for the same entity further contributes to attention drift, resulting in missed relational links.

In Figure 10b, we observe several examples of missed relationships. Numerical entities such as "1000," "10000," "7000," and "3000" are correctly linked to the main graph in LinkKG-no-str-prompt (Figure 10c), but are missing in LinkKG-no-coref (Figure 10b). Another notable observation is that these numerical entities are not extracted at all in LINK-KG, as they do not belong to any of the seven target entity types defined in the prompt. This demonstrates that coreference resolution and structured prompting complement each other in enabling more precise entity and relationship extraction.

b) Analysis of LinkKG-no-str-prompt Graph: The LinkKG-no-str-prompt graph performs better than LinkKG-no-coref in reducing duplication, primarily due to the presence of the coreference resolution module, which helps unify different surface forms of the same entity at an early stage. However, several duplicate entities still remain, such as "Tractor Trailer" and "Trailer", "Laredo" and "Laredo, Texas", "Yusif Ahmed Rufied" and "Defendant", as well as "Bruce Henderson" and "Agent Bruce Henderson". As discussed in Section V-A2, this residual duplication is largely caused by the use of the standard GraphRAG prompt. Due to its unconstrained extraction rules, the model tends to extract loosely related or irrelevant mentions, which further increases duplication.

Despite improved duplication control, LinkKG-no-str-prompt shows a noticeably higher level of noisy entity extraction. This is because it relies on the standard GraphRAG prompt, which lacks structured constraints for filtering irrelevant entities. As illustrated in Figure 10c, several unnecessary government-related entities are extracted around key nodes such as "Bruce Henderson" and "Yusif Ahmed Rufied." These noisy entities are highlighted using ovals in the graph. In addition, multiple fine-grained but irrelevant entities appear, including monetary mentions such as "$10000" and "$1000", as well as entities like "cab", "lady", and "rear of the tractor trailer". These are extracted due to the loose and unconstrained nature of the prompt. This highlights the importance of in-prompt filtering, sequential extraction, and well-defined entity types, which are central components of the LINK-KG structured prompting strategy.

At the same time, we observe that LinkKG-no-str-prompt captures certain fine-grained entities that provide useful operational context for analyzing human smuggling activities. For instance, details such as "Rear of the tractor trailer" where Martin Adaboh was seated, "$10000" as the amount he initially carried, "$1000" given to Yusif, and "$7000" as the remaining money offer insight into the financial and logistical aspects of the operation. These examples indicate that while aggressive filtering reduces noise, some fine-grained entities can still be valuable for investigative analysis. This emphasizes the need to balance noise control with the preservation of informative details.

One important observation is that in LinkKG-no-coref, due to the absence of the coreference resolution module, the "Defendant" node in Figure 10b exhibits a very high degree of connectivity. This occurs because the smuggler "Yusif Ahmed Rufied" is consistently referred to as "Defendant" throughout the document, and all such mentions

Category	TP	FP	FN	P	R	F1
Global Performance						
Entities	1066	392	470	0.7311	0.6940	0.7121
Relations	476	781	707	0.3787	0.4024	0.3902
Per-Type Entity Performance						
Person	406	148	262	0.7329	0.6078	0.6645
Means of Transportation	202	39	60	0.8382	0.7710	0.8032
Means of Communication	12	4	3	0.7500	0.8000	0.7742
Routes	60	7	4	0.8955	0.9375	0.9160
Location	275	153	96	0.6425	0.7412	0.6884
Smuggled Items	90	17	27	0.8411	0.7692	0.8036
Organization	11	3	14	0.7857	0.4400	0.5641

TABLE V: Global and per-type entity–relation extraction performance. TP, FP, and FN denote true positives, false positives, and false negatives, respectively. P, R, and F1 denote precision, recall, and F1-score.

are treated as separate surface forms of the same entity without being resolved. In contrast, in LinkKG-no-str-prompt (Figure 10c), which includes the coreference resolution module, these "Defendant" mentions are correctly unified and replaced with the actual smuggler name "Yusif Ahmed Rufied." As a result, the high-degree connectivity is transferred to the correct entity node. Although a residual "Defendant" node is still present in LinkKG-no-str-prompt, it has only limited connections, mainly due to partial replacements during resolution.

c) Analysis of LINK-KG Graph: Analyzing the LINK-KG graph reveals a well-connected structure with minimal duplication and noise, enabled by the integration of coreference resolution and structured prompts. No duplicate entities were observed in this case during manual inspection. Only one government-related node, "Immigration and Customs Enforcement," appears as a noisy extraction. This indicates that the cache-based coreference resolution strategy in LINK-KG is effective in consolidating entity mentions from noisy and unstructured legal text, while the structured prompts help reduce noise and enable precise entity and relationship extraction.

One remaining concern is the extraction of irrelevant placeholder nodes such as "No Specific Organization Mentioned," "No Specific Means of Communication Mentioned," and "No Specific Routes Mentioned." These appear to be artifacts of either LLM hallucination or forced extraction due to strict prompt rules. This suggests that further prompt refinement is needed to mitigate such cases. An alternative solution would be to introduce a lightweight post-extraction filtering layer to eliminate such irrelevant nodes.

Through this ablation study, we derive key insights into the impact of coreference resolution and structured prompting, along with their respective strengths and limitations. A key observation is that LINK-KG is able to construct a coherent knowledge graph directly from unstructured legal narratives. The resulting structured representation can effectively support investigative analysis and can be further extended to study the temporal evolution of networks, operational patterns, and related behavioral dynamics.

C. Entity and Relationship Evaluation

1) Global Entity and Relation Extraction Performance: Table V reports the global ER extraction performance across the entire evaluation dataset. For entity extraction, the model achieves an F1-score of 0.7121, with a precision of 0.7311 and a recall of 0.6940. This indicates that a large portion of the entities are correctly identified, as reflected by the high number of true positives (1066). The relatively balanced precision and recall show that the structured prompting strategy enables the model to extract entities with both reasonable accuracy and coverage.

The false positive count (392) is non-trivial, but our manual inspection reveals that a significant portion of these errors originate from limitations in the annotated dataset rather than incorrect model behavior. In several cases, the gold annotations omit certain valid entity mentions that are explicitly present in the text and correctly extracted by our prompt. As a result, these correct extractions are incorrectly counted as false positives during evaluation. This suggests that the reported precision is a conservative estimate and may underestimate the true extraction quality due to incomplete ground-truth labeling.

The false negative count (470) reflects missed entity extractions. These errors mainly occur in cases where multiple entities of the same type appear in a sentence and one of the less salient mentions is overlooked. For example, in one instance, locations such as "San Antonio, Texas" and "Laredo, Texas" are correctly extracted, but "Border

Patrol Checkpoint" is missed as a location entity. This likely happens because the surrounding context emphasizes city-level locations, which are more prominent than facility-type locations such as checkpoints. In another example, the term "smuggler" is missed as a *Person* entity. In that sentence, other person entities appear as proper names, while "smuggler" is expressed as a role rather than a named individual. As a result, the model may deprioritize it as an entity mention. These cases suggest that adding more role-based and facility-based examples into the prompt could further improve recall for such underrepresented patterns.

In contrast, relation extraction remains considerably more challenging. The model achieves a global relation F1-score of 0.3902, with a precision of 0.3787 and a recall of 0.4024. The large number of false positives (781) and false negatives (707) indicates that the model struggles to extract relationships reliably. One contributing factor is the limitation of the annotated dataset, where missing even a single entity can lead to multiple missing relationships involving that entity, thereby increasing both false positives and false negatives. Another important reason is the nature of legal narratives themselves. Legal documents are often dense, with many entities participating in multiple relationships within complex sentences. This makes accurate relationship extraction inherently difficult. These results highlight the challenges of relation extraction in unstructured legal text and motivate the need for improved extraction mechanisms. Further refinement of the annotated data is also expected to lead to more reliable relation extraction performance.

2) Per-Type Entity Analysis:

a) High-Performing Entity Types.: As shown in Table V, among all seven categories, *Routes* achieves the highest performance with an F1-score of 0.9160, supported by very high precision (0.8955) and recall (0.9375). This strong performance can be attributed to the limited variability in route mentions and their relatively consistent lexical patterns in legal text. It is also important to note that the total number of instances for the *Routes* category is only 64, suggesting that the observed high performance may be partly influenced by the limited data size.

Similarly, *Means of Transportation* and *Smuggled Items* also demonstrate strong performance, both achieving F1-scores above 0.80. This is particularly encouraging since the total number of entities for these two categories exceeds 100. These entity types typically appear in well-defined operational contexts (e.g., "truck," "trailer," "firearm," "currency"), which makes them easier to detect using structured prompts.

b) Moderate-Performing Entity Types.: The *Person* and *Location* categories show moderate performance, with F1-scores of 0.6645 and 0.6884, respectively. These two categories also have the largest number of total instances in the dataset (668 persons and 371 locations), which provides good variability but also makes them more prone to both false positives and false negatives. Person entities frequently appear under multiple surface forms (e.g., full names and roles such as "defendant"), which increases ambiguity. Location mentions often range from cities and states to fine-grained references such as checkpoints or districts, leading to missed or inconsistent extractions.

c) Low-Resource and Challenging Entity Types.: The weakest performance is observed for the *Organization* category, with an F1-score of 0.5641 and a particularly low recall of 0.44. This category also has one of the smallest sample sizes in the dataset (25 total instances), leading to strong class imbalance. Organizations in legal text often appear in abbreviated, implicit, or descriptive forms (e.g., "ICE," "BPA"), which makes them difficult to consistently normalize and extract. The *Means of Communication* category contains very few instances (15), yet it shows stable performance because these entities usually appear with clear and distinctive words such as "phone" and "WhatsApp," which makes them easier to detect.

One more important observation is that the model's pretrained knowledge also appears to influence extraction performance. Categories such as *Means of Communication*, *Routes*, and *Means of Transportation* often involve common terms like "phone," "WhatsApp," "Interstate 35," "truck," and "trailer," which frequently appear across many documents seen during model training. As a result, the model develops a stronger understanding of how these terms are used in context, which helps improve extraction accuracy.

In contrast, categories such as *Person* and *Organization* rely on specific names that can vary widely and may not be well represented in the training data, especially those related to human smuggling. This variability makes these entities harder to recognize consistently, which contributes to their relatively lower performance.

VI. Future Work

The ablation study highlights several limitations in the current LINK-KG pipeline, including the extraction of overly long entity spans and the presence of irrelevant entities such as "No Specific Organization Mentioned". These issues mainly arise due to LLM hallucinations and forced extraction behavior introduced by structured prompting.

This work can be extended by refining the prompts to enforce better control over the entity label length and to reduce forced placeholder outputs. In addition, a post-extraction filtering layer can be introduced to remove irrelevant or spurious entities. Such a filtering mechanism can also be useful for further reducing both node duplication and legal noise in the constructed knowledge graphs.

In addition, the framework can be extended to support coreference resolution across multiple documents, enabling the construction of coherent cross-document knowledge graphs. This extension would facilitate deeper analysis of inter-case linkages, temporal evolution, and broader operational patterns within human smuggling networks.

For ER evaluation, future work includes refining the annotated dataset to improve label consistency and coverage, which will lead to more accurate performance assessment. Another important direction is benchmarking against state-of-the-art ER systems to better position the performance of LINK-KG in a broader research context.

Another important step that is ongoing is incoporrateing the recnet LLMs in the ER extraction pipeline and check if there is an improvement since these modesl have been effectively trained and performing better in other benchmarks. Another ongoing effort is the integration of recently released open-source LLMs such as Qwen3 and GPT-OSS into the ER extraction pipeline. This work will provide a deeper understanding of how the current extraction pipeline performs relative to more recent models.

REFERENCES

[1] D. Meher, C. Domeniconi, and G. Correa-Cabrera, "Link-kg: Llm-driven coreference-resolved knowledge graphs for human smuggling networks," *arXiv preprint arXiv:2510.26486*, 2025.

[2] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitansky, R. O. Ness, and J. Larson, "From local to global: A graph rag approach to query-focused summarization," *arXiv preprint arXiv:2404.16130*, 2024.