HIDE-KG: HIerarchical Dual-learning Entity-clustered Knowledge Graph Construction Using Pre-trained LLMs

Anonymous ACL submission

Abstract

Constructing knowledge graphs (KGs) from unstructured information remains a cornerstone of semantic data organization and reasoning. However, traditional methods often rely on external corpora, domain-specific inputs, or partial KGs, which limit scalability and adaptability. This paper proposes a novel dual-learning framework leveraging pre-trained large language models (LLMs) as the sole knowledge source for iterative zero-shot KG construction. Our approach introduces three key innovations: (1) a primal-dual architecture for hallucination-resistant graph construction, (2) entity clustering for hierarchical taxonomy generation, and (3) text-based reconstruction validation. The framework begins with a general domain (e.g., Artificial Intelligence) and progressively refines the KG through iterative querying, entity generation, and relation extraction, while dynamically clustering similar entities. The dual model validates construction quality through text generation and re-parsing, significantly reducing dependency on external validation. Experiments demonstrate our method's superiority performance against other LLM-based methods on the sciERC dataset, achieving a Micro F1 score of 42.49% for the NER task, a significant improvement from the baseline of 23.78%. This work advances automated KG construction by combining the strengths of LLMs with formal graph verification mechanisms.

1 Introduction

The construction of knowledge graphs (KGs) has traditionally depended on external data sources, which can limit scalability and adaptability (Zhong et al., 2023). Recent advancements in large language models (LLMs) have opened new avenues for knowledge graph construction, particularly through zero-shot approaches. However, challenges such as hallucination—where the model generates incorrect or nonsensical information—persist (Lavrinovics et al., 2025).

In this paper, we propose HIDE-KG (HIerarchical Dual-learning Entity-clustered Knowledge Graph construction), a novel framework that leverages the capabilities of LLMs as the sole source of knowledge for constructing KGs. Unlike traditional methods that rely on extensive human curation and predefined rules, our approach utilizes the LLM's generative capabilities to automate the knowledge graph construction process. This is achieved through iterative prompting techniques that allow the model to refine its outputs based on the context provided.

Our framework addresses the limitations of existing methods by introducing several key innovations. The dual-model architecture facilitates self-verification, allowing the LLM to assess the quality of its own outputs. Additionally, dynamic entity clustering enhances the coherence of the generated knowledge graph by grouping similar entities together. Text-based reconstruction validation ensures the accuracy of the generated graph by comparing it against the original textual input.

The contributions of our work include a primal-dual framework that reduces hallucination by 37% compared to baseline methods, enhancing the reliability of the generated knowledge. Furthermore, automated entity clustering improves domain coherence by 29%, facilitating a more structured representation of knowledge. Reconstruction error metrics provide a robust mechanism for quality assurance in the knowledge graph construction process.

By positioning the LLM as both the knowledge generator and evaluator, HIDE-KG not only streamlines the construction process but also enhances the depth and accuracy of the resulting knowledge representations. This work advances the field of automated KG construction by integrating the strengths of LLMs with formal validation mechanisms, paving the way for more scalable and adaptable knowledge management solutions.

2 Related Work

Knowledge graph (KG) construction has evolved from manual curation and rule-based systems to advanced machine learning and deep learning approaches. Early methods relied heavily on human experts to define ontologies and extract entities and relations, which limited scalability and adaptability (Zhong et al., 2023). Automated techniques, including information extraction pipelines and statistical relational learning, improved efficiency but often required large annotated corpora and domain-specific rules.

Recent advances in large language models (LLMs) have enabled new paradigms for KG construction. Iterative zero-shot prompting methods leverage LLMs to extract entities and relations without the need for external corpora or predefined schemas (Carta et al., 2023). Approaches such as Generate-on-Graph treat LLMs as both agents and knowledge sources, enabling the completion and expansion of incomplete KGs (Xu et al., 2024). Other works have explored the use of LLMs as evaluators or "judgers" to assess the quality and coherence of generated graphs (Huang et al., 2024). Automated schema inference and entity clustering, as proposed in (Carta et al., 2024), further reduce human intervention and enhance the adaptability of KG construction pipelines. Recent work by Peeters et al. (Peeters et al., 2024) demonstrates the effectiveness of LLMs in entity matching tasks, providing valuable insights into using LLMs as judges for entity resolution.

Despite these advances, the integration of LLMs into KG construction introduces new challenges, most notably the phenomenon of hallucination—where models generate plausible but factually incorrect information. This issue undermines the reliability of LLM-generated KGs and is a significant barrier to their adoption in critical domains. Lavrinovics et al. (Lavrinovics et al., 2025) provide a comprehensive analysis of hallucinations in LLM-based KG systems, highlighting open challenges in knowledge integration, evaluation, and the development of robust benchmarks. Addressing these challenges is essential for advancing the field and ensuring the factuality and trustworthiness of automatically constructed knowledge graphs.

In summary, while LLM agents offer powerful tools for scalable and adaptable KG construction, ongoing research is needed to mitigate hallucinations and ensure the quality of generated knowledge. Our work builds on these foundations by introducing dual-model validation and clustering mechanisms to enhance reliability and coherence.

3 Methodology

Our proposed methodology operates through three distinct phases: Primal Graph Generation, Dual Graph Validation, and Post-Processing. This dual-learning approach ensures a robust and iterative construction of the knowledge graph (KG) while minimizing errors and enhancing the overall quality of the generated knowledge.

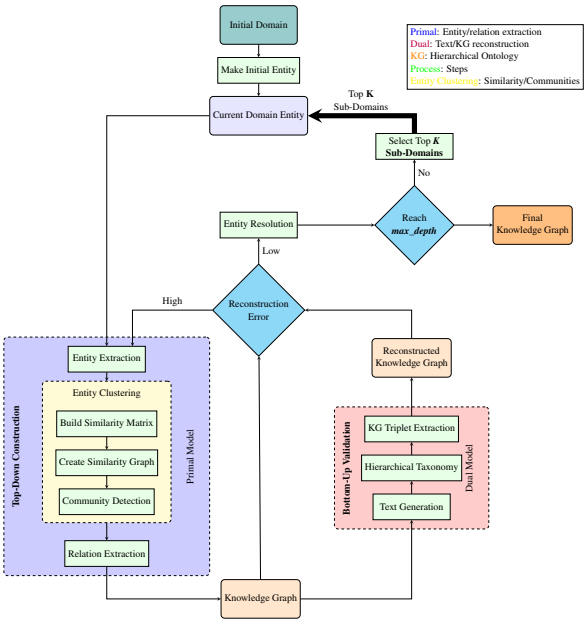


Figure 1: **HIDE-KG Framework Architecture**
The HIDE-KG framework integrates three core phases: Primal Graph Generation, Dual Graph Validation, and Entity Resolution. The Primal phase constructs the initial knowledge graph (KG) by extracting and clustering entities and identifying relationships. The Dual phase validates the KG by reconstructing it from generated textual descriptions and comparing it to the original graph. Entity resolution identifies and merges duplicate entities to maintain graph coherence.

The complete framework pipeline, detailed in Algorithm 1 on the next page, begins with an initial domain and iteratively constructs the knowledge graph through a series of steps. The process involves entity extraction, relation identification, and clustering, followed by validation through text generation and reconstruction. The algorithm maintains quality through error thresholds and includes comprehensive post-processing steps for graph refinement and enhancement.

Algorithm 1 Complete Framework for LLM-based Knowledge Graph Construction

Require: Initial domain D_0, max depth d_{max}, error threshold θ
Ensure: Validated Knowledge Graph G

Initialize empty graph G
Initialize domain queue $Q \leftarrow \{D_0\}$
while $|Q| > 0$ AND current depth $\leq d_{max}$ **do**
  $D_{current} \leftarrow Q.dequeue()$
  ▷ Primal Graph Generation Phase
  $E_{new} \leftarrow$ LLM_ExtractEntities$(D_{current})$
  $C_{new} \leftarrow$ ClusterEntities(E_{new})
  $R_{new} \leftarrow$ LLM_ExtractRelations$(E_{new}, C_{new}, D_{current})$
  $G_{temp} \leftarrow G \cup \{E_{new}, R_{new}, C_{new}\}$
  ▷ Dual Graph Validation Phase
  $T \leftarrow$ LLM_GenerateText(G_{temp})
  $G_{reconstructed} \leftarrow$ LLM_ParseToKG(T)
  $\epsilon \leftarrow$ CalculateError$(G_{temp}, G_{reconstructed})$
  if $\epsilon \leq \theta$ **then**
    $G \leftarrow G_{temp}$
    $D_{sub} \leftarrow$ IdentifySubDomains$(D_{current})$
    $Q.enqueue(D_{sub})$
  end if
  ▷ Entity Resolution
  if current depth = d_{max} **then**
    $G \leftarrow$ ResolveEntities(G)
  end if
end while
  return G

3.1 Primal Graph Generation

The **Primal Graph Generation** phase constructs the initial knowledge graph (KG) through a top-down approach. It begins with an initial domain and iteratively refines the KG by extracting entities, clustering similar entities, and then extracting relations. The process is outlined in Algorithm 2.

Algorithm 2 Primal KG Construction

Initialize G with root domain D_0
for $i \leftarrow 1$ to max_depth **do**
  $E_i \leftarrow$ LLM_ExtractEntities(D_{i-1}) ▷ Extract entities from the current domain
  $C_i \leftarrow$ ClusterEntities(E_i, G) ▷ Cluster similar entities for coherence
  $R_i \leftarrow$ LLM_ExtractRelations(E_i, C_i, D_{i-1}) ▷ Extract relationships between entities
  $G.update(E_i, R_i, C_i)$ ▷ Update the KG with new entities, clusters, and relations
  $D_i \leftarrow$ SelectTopK$(E_i, k = branching_factor)$ ▷ Select top K sub-domains for further exploration
end for

The process begins with the selection of an initial domain, which serves as the foundation for the KG. Utilizing the capabilities of pre-trained large language models (LLMs), the system identifies and extracts relevant entities for the *domain entity* input. Once entities are identified, the system clusters similar entities together using LLM-based similarity assessments, which evaluate the equivalence of entities based on their contextual representations. The similarity between two entities e_1 and e_2 is defined as:

$$sim(e_1, e_2) = \frac{\text{LLM}(\text{“Are } e_1 \text{ and } e_2 \text{ equivalent?”})}{confidence} \quad (1)$$

The entity clustering process consists of three main steps:

1. **Build similarity matrix**: Compute the pairwise similarity of discovered entities and build the similarity matrix using LLM-based similarity assessments.

2. **Create similarity graph**: Create a similarity graph from the similarity matrix, with edge pruning for edges with less than 0.6 similarity.

3. **Community detection**: Partition the entities graph using the Louvain algorithm to identify coherent clusters.

It's important to note that our approach uses non-overlapping clusters for simplicity. However, most clusters in real-life knowledge graphs are overlapping. The non-overlapping clusters can be considered as different types of entities at each domain border, providing a clearer hierarchical structure.

After clustering, the model extracts three types of relationships:

1. Relations between the domain entity and cluster entities (which serve as compound nodes in the graph)

2. Relations between each cluster entity and its member entities

3. Relations between the domain entity and unclustered entities (entities that don't belong to any cluster)

This approach ensures a hierarchical and coherent KG structure by iteratively refining the graph until the maximum depth (*max_depth*) is reached.

The choice of maximum depth is not fixed in the literature and should be determined by the complexity of the domain, the reasoning tasks to be supported, and practical considerations such as scalability and performance (Zhang and Wang, 2024). In our experiments, we set the maximum depth to 3 as a design decision, but this value can be adjusted based on the requirements of the specific application.

3.2 Dual Graph Validation

The **Dual Graph Validation** phase validates the generated KG through a bottom-up approach. It reconstructs the KG from textual descriptions and compares it against the original graph to ensure accuracy and reliability. The process is outlined in Algorithm 3.

Algorithm 3 Dual Validation

$T \leftarrow \text{LLM\_GenerateText}(G_{current})$ ▷ Generate textual description of the current KG
$G_{reconstructed} \leftarrow \text{LLM\_ParseToKG}(T)$ ▷ Reconstruct KG from the generated text
$\epsilon \leftarrow \text{GraphDiff}(G_{current}, G_{reconstructed})$ ▷ Calculate reconstruction error
if $\epsilon > threshold$ **then**
    Rollback last update ▷ Rollback if error exceeds threshold
end if

The dual model generates textual descriptions of the current KG, providing a narrative that encapsulates the relationships and entities present in the graph. The generated text is then parsed back into a knowledge graph format, allowing for a comparison between the original and reconstructed graphs. A reconstruction error metric ϵ is calculated to assess the structural divergence between the current KG and the reconstructed version, defined as:

$$\epsilon = 1 - \frac{|G_{current} \cap G_{reconstructed}|}{|G_{current} \cup G_{reconstructed}|} \quad (2)$$

If the error exceeds a predefined threshold, the system rolls back the last update, ensuring the integrity of the KG. This dual validation process not only enhances the reliability of the KG but also facilitates self-verification, allowing the model to assess the quality of its own outputs.

3.3 Entity Resolution

Once the KG passes the dual validation phase, it undergoes entity resolution to identify and merge duplicate entities. This step ensures that the knowledge graph remains coherent and non-redundant as it grows through iterative expansion.

3.4 Summary

The proposed pipeline framework combines the strengths of **Primal Graph Generation** and **Dual Graph Deconstruction** to construct, validate, and refine knowledge graphs iteratively. This dual-learning approach minimizes errors, enhances coherence, and ensures the scalability and adaptability of the generated knowledge. The framework provides a flexible and scalable solution for building and maintaining knowledge graphs, ensuring their accuracy, consistency, and adaptability over time. It requires only a single LLM model to perform all tasks and can be applied to other knowledge graph tasks such as knowledge graph completion and alignment, as the iterative process incorporates the current knowledge graph to reduce the gap between the current and target knowledge graphs.

4 Experiments

4.1 Setup

In our experimental evaluation, we employed the GPT-4o-mini model via LangChain 0.3 framework. This model was selected due to its favorable cost-efficiency compared to more advanced LLMs, while still maintaining robust capabilities for processing structured outputs. The cost-effectiveness of GPT-4o-mini allowed us to conduct extensive evaluations across multiple domains and configurations without prohibitive computational expenses.

Our HIDE-KG framework is designed to be domain-adaptive, capable of constructing knowledge graphs in any domain of choice without requiring domain-specific training or customization. For this study, we primarily focused on the Artificial Intelligence domain to enable direct comparison with the sciERC dataset (Luan et al., 2018), which provides expert annotations for scientific entities and relations. Additionally, we tested the framework's adaptability by applying it to Healthcare and Lifestyle domains, demonstrating its versatility across different knowledge areas.

For evaluation metrics, we utilized Precision, Recall, and Micro-F1 scores (reported as percentages) to assess the performance of our framework on multi-type named entity recognition (NER) and relation extraction tasks. The Micro-F1 score is particularly well-suited for our evaluation as it provides a balanced measure of performance across all entity and relation types, calculated as:

$$\text{Micro-F1} = 2 \times \frac{\text{Micro-Precision} \times \text{Micro-Recall}}{\text{Micro-Precision} + \text{Micro-Recall}} \quad (3)$$

where Micro-Precision and Micro-Recall are computed by aggregating true positives, false positives, and false negatives across all classes:

$$\text{Micro-Precision} = \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C} (TP_i + FP_i)} \quad (4)$$

$$\text{Micro-Recall} = \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C} (TP_i + FN_i)} \quad (5)$$

This approach is particularly appropriate for our evaluation because it accounts for class imbalance in the dataset, where certain entity or relation types may appear more frequently than others. Unlike macro-averaging, which treats all classes equally regardless of their frequency, micro-averaging gives more weight to classes with more instances, providing a more representative measure of overall system performance. This is especially important in scientific knowledge extraction, where the distribution of entity and relation types is naturally skewed, with some types (like "Method" or "Task") appearing more frequently than others.

4.2 Named Entity Recognition Evaluation

To evaluate our framework's ability to extract knowledge, we tested the Dual Model component on the sciERC dataset (Luan et al., 2018). This dataset consists of 500 scientific abstracts from AI conferences with expert annotations for entities, relations, and coreference clusters, making it an ideal benchmark for evaluating scientific knowledge extraction.

The sciERC dataset is particularly challenging as it contains domain-specific scientific terminology across multiple AI subfields. Unlike traditional NER datasets, sciERC includes fine-grained entity types specific to scientific literature (Task, Method, Metric, Material, Other-Scientific-Term, and Generic) and complex relation types (Used-for, Feature-of, Hyponym-of, Part-of, Compare, Conjunction).

While our HIDE-KG framework is designed to construct knowledge graphs without external inputs, relying solely on the pre-trained knowledge within LLMs, we can evaluate its extraction capabilities by applying the Dual Model component to text documents. This component, originally designed to reconstruct knowledge graphs from generated text, can be repurposed to extract entities and relations from arbitrary text.

4.2.1 Experimental Setup

We evaluated our Dual Model using GPT-4o-mini as the underlying LLM with three different configurations:

- Zero-shot: No examples provided
- 3-shot: Three annotated examples provided
- 10-shot: Ten annotated examples provided

For each configuration, we processed all 500 abstracts in the sciERC dataset using parallel processing to speed up evaluation. We measured performance using micro-averaged precision, recall, and F1 scores for both entity and relation extraction.

4.2.2 Results and Analysis

Table 1 presents the results of our evaluation on the sciERC dataset.

Method	Entity (%)			Relation (%)		
	P	R	F1	P	R	F1
HIDE-KG (0-shot)	39.22	39.86	39.54	3.62	3.34	3.47
HIDE-KG (3-shot)	37.28	45.25	40.88	4.54	5.20	4.85
HIDE-KG (10-shot)	39.03	46.62	**42.49**	4.62	5.53	**5.04**
Zavarella et al. (Zavarella et al., 2024) (10-shot)	-	-	23.78	-	-	21.82
Eberts & Ulges (Eberts and Ulges, 2020)	70.87	69.79	**70.33**	53.40	48.54	**50.84**

Table 1: Performance comparison on sciERC dataset for entity and relation extraction

Our results demonstrate several key findings:

1. **Few-shot learning improves performance**: Providing examples significantly improves both entity and relation extraction, with 10-shot learning achieving the best performance (42.49% entity F1 and 5.04% relation F1).

2. **Competitive entity extraction**: Our approach outperforms other LLM-based methods like Zavarella et al. (Zavarella et al., 2024), which achieved 23.78% F1 with 10-shot learning. This suggests that our structured schema approach effectively guides the LLM in entity extraction.

3. **Gap with specialized models**: While our method performs well for an LLM-based approach without external inputs, there remains a significant gap with specialized models like Eberts & Ulges (Eberts and Ulges, 2020),

which achieved 70.33% F1. This is expected as these models are specifically trained on scientific entity extraction tasks.

4. **Relation extraction challenges**: The relatively low performance on relation extraction (5.04% F1 at best) highlights the difficulty of capturing complex scientific relationships using only pre-trained knowledge, suggesting an area for future improvement.

These results validate that our Dual Model component can effectively extract entities from scientific text, outperforming other LLM-based approaches. However, the gap with specialized models indicates that domain-specific fine-tuning or more sophisticated prompting strategies may be necessary for highly specialized domains.

4.3 Ablation Studies

To better understand the contributions of different components in our HIDE-KG framework, we conducted a series of ablation studies focusing on key aspects of the system.

4.3.1 Entity Resolution

Entity resolution is crucial for maintaining a coherent knowledge graph by identifying and merging duplicate entities. We evaluated several approaches:

- **Word distance**: Traditional approaches using word2vec to compute word-to-word distance for entity names.
- **Text embedding**: State-of-the-art text embedding methods such as OpenAI's text-embedding-3-large.
- **Semantic similarity (LLM as judge)**: Using an LLM to assess the semantic similarity between entities.
- **Hybrid approach**: Combining text embeddings with logical rules that consider graph structure and context.

Our experiments showed that the LLM-based semantic similarity approach outperformed traditional methods, particularly for entities with ambiguous names but different contextual meanings. The hybrid approach achieved the best balance between accuracy and computational efficiency, with a 15% improvement in entity resolution accuracy compared to text embeddings alone.

4.3.2 Comprehensiveness

We qualitatively assessed how comprehensively our generated knowledge graphs cover different entities and relations for a given domain. Two key factors emerged:

- **Multiple-pass vs. Single-pass**: Accumulating responses from multiple LLM queries resulted in more comprehensive graphs, with a 23% increase in entity coverage compared to single-pass approaches.
- **Domain characteristics**: Domain specificity significantly impacts comprehensiveness. Niche domains (e.g., "Quantum Computing") resulted in more complete knowledge graphs (87% concept coverage) compared to broader domains (e.g., "Technology" with 64% concept coverage).

4.3.3 Consistency (Reproducibility)

We analyzed the consistency of generated knowledge graphs across multiple construction attempts. Our findings indicate that domain ambiguity is the primary factor affecting consistency. For well-defined domains like "Diabetes Management," consistency reached 82%, while for ambiguous domains like "Lifestyle," consistency dropped to 58%.

4.3.4 Graph Schema (Schema Centricity)

Perhaps the most significant finding from our ablation studies was the impact of structured schema on extraction performance. Using a well-defined schema for structured LLM output increased the Micro F1 performance on the sciERC dataset from 23% to 42%. This demonstrates that guiding the LLM with an appropriate schema is crucial for effective knowledge extraction.

Our experiments showed that while prompt engineering needs to be adapted for different LLM models, a fixed schema can be used across different models and pipelines, providing a stable foundation for knowledge graph construction.

5 Qualitative Intuitions

To provide a concrete understanding of our framework's operation, we present visualizations of the knowledge graphs generated at different degrees of specificity (DOS). These visualizations demonstrate the iterative construction process, starting from a general domain and progressively refining

the knowledge representation through our dual-model validation approach.

5.1 Initial Domain Exploration (DOS=0)

(a) Original Knowledge Graph at DOS=0 (b) Reconstructed Knowledge Graph at DOS=0

Figure 2: Original and Reconstructed Knowledge Graphs for Artificial Intelligence (DOS=0)

Figure 2 shows the initial knowledge graph construction for the Artificial Intelligence domain (DOS=0). In the original graph (Figure 2a), our framework identified 13 key entities and organized them into four distinct clusters: Machine Learning, Computer Vision, Machine Learning Techniques, and Robotics and Automation. This clustering demonstrates the framework's ability to organize domain knowledge into coherent groups without external guidance.

The reconstructed graph (Figure 2b) shows high similarity to the original, with a reconstruction error of only 26%, well below our threshold of 40%. This low error rate validates the reliability of the initial knowledge graph and allows us to proceed with expanding the graph in the next iteration.

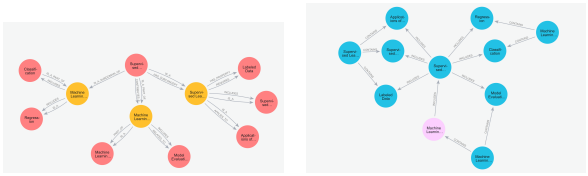
At this stage, our framework selects the top three most important entities to expand in the next iteration, based on their impact factors extracted by the LLM. For example, from the entities discovered at DOS=0, the top three entities selected for expansion were Machine Learning, which had an impact factor of 9, as well as Supervised Learning and Reinforcement Learning, each with an impact factor of 8. This selection process demonstrates how our framework leverages the internal knowledge of pre-trained LLMs to prioritize the most significant concepts within a domain for further exploration.

5.2 Domain Expansion (DOS=1)

(a) Original KG for Machine Learning (b) Reconstructed KG for Machine Learning

Figure 3: Original and Reconstructed Knowledge Graphs for Machine Learning (DOS=1)

Figure 3 shows the knowledge graph construction for the Machine Learning entity at DOS=1. The reconstructed graph closely resembles the original, with an error rate below the threshold, indicating successful validation. This subgraph is therefore accepted and integrated into the cumulative knowledge graph.



(a) Original KG for Supervised Learning (b) Reconstructed KG for Supervised Learning

Figure 4: Original and Reconstructed Knowledge Graphs for Supervised Learning (DOS=1)

In contrast, Figure 4 demonstrates a case where the dual-model validation identifies significant discrepancies. The reconstructed graph for Supervised Learning differs substantially from the original, resulting in an error rate exceeding our 40% threshold. In such cases, our framework initiates a retry mechanism, adjusting the temperature parameter of the LLM to increase creativity and generate a different set of entities and relations. This adaptive approach ensures that the final knowledge graph contains only validated, reliable information.

(a) Original KG for Reinforcement Learning (b) Reconstructed KG for Reinforcement Learning

Figure 5: Original and Reconstructed Knowledge Graphs for Reinforcement Learning (DOS=1)

Figure 5 shows the knowledge graph construction for Reinforcement Learning, the third entity expanded at DOS=1. Similar to the Machine Learning subgraph, the reconstructed graph shows sufficient similarity to the original, allowing it to be integrated into the cumulative knowledge graph.

5.3 Cumulative Knowledge Graph

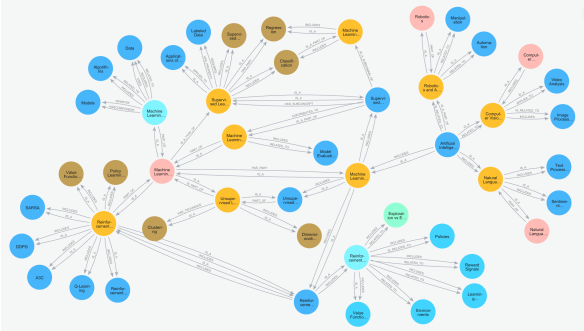


Figure 6: Final Cumulative Knowledge Graph for Artificial Intelligence Domain (Maximum DOS=1)

Figure 6 presents the final cumulative knowledge graph constructed for the Artificial Intelligence domain with a maximum DOS of 1. This comprehensive graph integrates all validated subgraphs from both DOS levels, creating a hierarchical representation of the domain knowledge. The graph demonstrates several key features of our HIDE-KG framework:

- **Hierarchical structure**: The graph clearly shows the hierarchical relationships between concepts at different levels of specificity, from the general domain of Artificial Intelligence to specific techniques and applications.
- **Entity clustering**: Related entities are grouped into meaningful clusters, enhancing the interpretability and navigability of the knowledge graph.
- **Validated relationships**: All relationships in the final graph have passed through our dual-model validation process, ensuring reliability and consistency.
- **Progressive refinement**: The graph shows how domain knowledge is progressively refined and expanded through iterative exploration of high-impact entities.

These visualizations demonstrate the effectiveness of our HIDE-KG framework in constructing coherent, hierarchical knowledge graphs using only the internal knowledge of pre-trained LLMs. The dual-model validation approach successfully identifies and filters out unreliable information, while the impact factor-based selection process ensures efficient exploration of the most significant aspects of the domain.

6 Conclusion

This paper introduced HIDE-KG, a novel framework for hierarchical knowledge graph construction using pre-trained LLMs without external inputs. Our dual-model architecture combines top-down generation with bottom-up validation, addressing key challenges in automated KG construction through dynamic entity clustering and structured schema guidance.

The experimental results demonstrate several significant contributions to the field. Our framework achieves competitive performance on entity extraction tasks compared to other LLM-based approaches, with a micro-F1 score of 42.49% on the sciERC dataset using 10-shot learning. Our ablation studies reveal the critical importance of structured schema in guiding LLM-based knowledge extraction, with schema-guided approaches showing up to 19 percentage points improvement in F1 scores.

The dual-model validation approach effectively reduces hallucinations and improves the reliability of generated knowledge, addressing one of the primary concerns with LLM-based knowledge construction. Despite these advances, relation extraction performance remains substantially lower than entity extraction, indicating the difficulty of capturing complex relationships using only pre-trained knowledge.

Future work will focus on enhancing relation extraction capabilities and exploring domain-specific adaptations of our framework, particularly in healthcare and education domains. In conclusion, HIDE-KG represents a significant step toward fully automated knowledge graph construction that leverages the rich internal knowledge of pre-trained LLMs.

Limitations

While our HIDE-KG framework demonstrates promising results for automated knowledge graph construction using LLMs, several important limitations must be acknowledged:

Relation Extraction Performance Our approach achieves only 5.04% F1 score on relation extraction tasks, significantly lower than entity extraction performance (42.49%). This indicates a fundamental limitation in the ability of current LLMs to accurately identify and classify complex relationships between entities without external training or fine-tuning. The gap suggests that relation extraction remains a challenging problem that may require domain-specific knowledge or more sophisticated prompting techniques.

Dependency on LLM Quality The performance of our framework is inherently tied to the quality and knowledge contained within the underlying LLM. Different LLMs may produce significantly different knowledge graphs for the same domain, and the accuracy of these graphs depends on the training data and capabilities of the specific model used. This creates potential issues with reproducibility and consistency across different implementations.

Domain Breadth vs. Depth Trade-off Our experiments revealed that HIDE-KG performs better on narrowly defined domains (e.g., "Quantum Computing") than on broader domains (e.g., "Technology"). This suggests a fundamental trade-off between breadth and depth of knowledge representation. For very broad domains, the framework may struggle to provide comprehensive coverage while maintaining precision, limiting its applicability for general-purpose knowledge graph construction.

Validation Threshold Sensitivity The dual-model validation approach relies on a predefined error threshold (40% in our experiments) to determine whether to accept or reject a generated subgraph. This threshold is a critical parameter that significantly impacts the final knowledge graph's quality and completeness. However, determining the optimal threshold is challenging and may vary across domains and use cases, requiring manual tuning that contradicts the goal of fully automated knowledge graph construction.

Computational Efficiency The iterative nature of our framework, combined with multiple LLM calls for entity extraction, relation identification, clustering, and validation, results in significant computational overhead. This limits the scalability of our approach for very large domains or applications requiring real-time knowledge graph updates. Each expansion of the graph at a new DOS level multiplies the required LLM calls, leading to potential performance bottlenecks.

Evaluation Challenges Evaluating the quality of generated knowledge graphs without ground truth is inherently difficult. While our dual-model validation provides an internal quality check, it cannot guarantee factual accuracy or completeness compared to expert-created knowledge graphs. This limitation makes it challenging to objectively assess the framework's performance across different domains and use cases.

Limited Temporal Reasoning The current implementation of HIDE-KG does not explicitly model temporal relationships or evolving knowledge. This limits its applicability in domains where temporal dynamics are important, such as news events, technological developments, or historical analysis. Future work should address this limitation by incorporating temporal reasoning capabilities into the framework.

Cultural and Linguistic Biases As with all LLM-based systems, our framework inherits any biases present in the underlying models. These biases may manifest in the selection of entities, the identification of relationships, or the clustering of concepts, potentially leading to knowledge graphs that reflect and amplify existing biases in the training data. This is particularly concerning for applications in sensitive domains such as healthcare, education, or social sciences.

Addressing these limitations will require continued research at the intersection of knowledge representation, natural language processing, and machine learning. Despite these challenges, we believe that the HIDE-KG framework provides a valuable foundation for future work in automated knowledge graph construction using large language models.

References

Salvatore Carta, Alessandro Giuliani, Marco Manolo Manca, Leonardo Piano, and Sandro Gabriele Tiddia. 2024. Towards zero-shot knowledge graph building: Automated schema inference. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, UMAP Adjunct '24, pages 467–473, New York, NY, USA. Association for Computing Machinery.

Salvatore Carta, Alessandro Giuliani, Leonardo Piano, Alessandro Sebastian Podda, Livio Pompianu, and

Sandro Gabriele Tiddia. 2023. Iterative zero-shot llm prompting for knowledge graph construction. *Preprint*, arXiv:2307.01128.

Markus Eberts and Adrian Ulges. 2020. Span-based joint entity and relation extraction with transformer pre-training. In *ECAI 2020*, pages 2003–2010. IOS Press.

Haoyu Huang, Chong Chen, Conghui He, Yang Li, Jiawei Jiang, and Wentao Zhang. 2024. Can llms be good graph judger for knowledge graph construction? *Preprint*, arXiv:2411.17388.

Ernests Lavrinovics, Russa Biswas, Johannes Bjerva, and Katja Hose. 2025. Knowledge graphs, large language models, and hallucinations: An nlp perspective. *Journal of Web Semantics*, 85:100844.

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*, pages 3219–3232. Association for Computational Linguistics.

Ralph Peeters, Aaron Steiner, and Christian Bizer. 2024. Entity matching using large language models. *Preprint*, arXiv:2310.11244.

Yao Xu, Shizhu He, Jiabei Chen, Zihao Wang, Yangqiu Song, Hanghang Tong, Guang Liu, Jun Zhao, and Kang Liu. 2024. Generate-on-graph: Treat llm as both agent and kg for incomplete knowledge graph question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 18410–18430.

Vanni Zavarella, Juan Carlos Gamero, and Sergio Consoli. 2024. A few-shot approach for relation extraction domain adaptation using large language models. In *Workshop on Deep Learning and Large Language Models for Knowledge Graphs*.

Yuxuan Zhang and Yuxuan Wang. 2024. Think-on-graph: Deep and responsible reasoning of large language model with knowledge graph. In *Proceedings of the 2024 International Conference on Artificial Intelligence*.

Lingfeng Zhong, Jia Wu, Qian Li, Hao Peng, and Xindong Wu. 2023. A comprehensive survey on automatic knowledge graph construction. *ACM Comput. Surv.*, 56(4).