CogCommon: Enhancing Cross-Domain Knowledge Extraction with LLM-Assisted Commonality Discovery

Chong Hu[1] Hangwei Qian[2] Ivor Tsang[2]

[1]*Nanyang Technological University, Singapore* [2]*CFAR and IHPC, Agency for Science, Technology and Research (A*STAR), Singapore*. Correspondence to: Hangwei Qian qian_hangwei@cfar.a-star.edu.sg.

1. Introduction

Knowledge fragmentation across specialized scientific subfields creates "knowledge islands" that hinder cross-domain integration [1, 2]. While this enables deep expertise, the lack of cross-domain commonality recognition in specific fields leads to inaccurate relationship identification across domains. For instance, in the semiconductor field, *Beam Steering* in FlatOptics and *Directional Beamforming* in mmWave share similar wavefront control principles but remain disconnected in traditional knowledge graphs. Similarly, *Waveguide* in SiPhotonics is isolated despite its relevance to beamforming techniques.

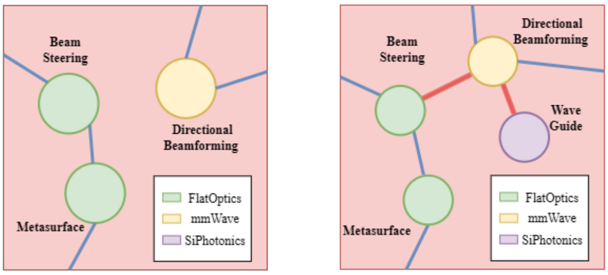

Fig. 1: Comparisons of entity relationships before and after applying our method. Our method introduces **cross-domain links** (shown in red), enabled by commonality integration.

To address this, we propose **CogCommon**, a framework that leverages cognitive linguistics to identify and integrate deep cross-domain structures. Our contributions include: 1) a two-stage method for discovering and incorporating cross-domain patterns into knowledge graphs, 2) a schema representation Ω capturing shared cognitive structures, and 3) experiments in **semiconductor** domains (FlatOptics, mmWave, SiPhotonics), demonstrating improved entity extraction and cross-domain connectivity. **Our results confirm the presence of cross-domain commonalities in semiconductor fields and show that leveraging them enhances entity extraction and knowledge integration.**

2. The Proposed CogCommon Method

Our approach, CogCommon, draws on Lakoff and Johnson's Conceptual Metaphor Theory [3], which establishes that human conceptual systems are fundamentally organized through shared cognitive structures. For example, our analysis identifies the conceptual metaphor "Wave Manipulation as Navigation System" across electromagnetic domains, where both optical metasurfaces and mmWave antennas conceptualize beam control through the same cognitive structure despite operating at vastly different wavelengths.

Given a set of domains $\mathcal{D} = \{D_1, D_2, ..., D_n\}$ with corresponding document sets, we define the commonality discovery problem as identifying a rich set of shared cognitive elements: $\Omega = \{\mathcal{M}, \mathcal{S}, \mathcal{A}, \mathcal{C}\}$ where $\mathcal{M}$ represents conceptual metaphors, $\mathcal{S}$ denotes structural patterns, $\mathcal{A}$ identifies abstraction hierarchies, and $\mathcal{C}$ encompasses cognitive frames.

Stage 1: Deep Commonality Discovery

Our framework implements a commonality extraction pipeline:

Domain-Specific Analysis: Given a domain-specific document corpus $\mathcal{D}$, domain-relevant knowledge elements are extracted from each document $D_i \in \mathcal{D}$:

$$D_i \overset{\mathcal{F}_D}{\longmapsto} \{E_i, R_i, C_i\},$$

where $\mathcal{F}_D$ represents the domain knowledge extraction process that identifies commonalities across the corpus. E_i, R_i, and C_i represent the sets of extracted entities, relations, and concepts, respectively.

Cross-Domain Pattern Recognition: LLMs are guided to identify shared cognitive structures across domains, uncovering common patterns that bridge domain differences. For details on the prompting strategy, refer to Appendix E.

Stage 2: Commonality-Guided Knowledge Graph Construction The extracted commonalities guide knowledge graph construction through our dual-alignment mechanism:

Commonality-Enhanced Extraction: For every text chunk c, extraction is guided by the identified commonalities:

$$c, \Omega \overset{\mathcal{F}_E}{\longmapsto} \{(e, t, m) \mid e \in c, t \in \mathcal{T}, m \in \mathcal{M}\},$$

$$c, \Omega \overset{\mathcal{F}_R}{\longmapsto} \{(e_1, r, e_2, s) \mid e_1, e_2 \in c, r \in \mathcal{R}, s \in \mathcal{S}\},$$

where $\mathcal{F}_E$ and $\mathcal{F}_R$ are entity and relation extraction functions that leverage commonalities to enhance cross-domain knowledge extraction, $\mathcal{T}$ is the set of entity types, and $\mathcal{R}$ is the set of relation types.

Cognitive Alignment: Entities and relationships are aligned to shared cognitive structures, creating a knowledge representation that captures deep commonalities across domains.

Table 1: Entity extraction performance comparison of different models across three different domains.

Model	Method	FlatOptics		mmWave		SiPhotonics		Average	
		Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)
gpt-4o-mini	GraphRAG	5.41	8.42	13.46	19.44	9.84	13.19	9.57	13.68
	CogCommon	**14.55**	**21.05**	10.64	14.93	**17.19**	**23.4**	**14.13**	**19.79**
qwen-plus	GraphRAG	7.14	11.76	8.00	11.43	10.00	15.38	8.38	12.86
	CogCommon	**11.43**	**19.05**	**11.76**	**19.67**	**16.10**	**25.68**	**13.10**	**21.47**
llama3.3-70b-instruct	GraphRAG	9.09	13.16	9.76	13.11	7.69	10.53	8.85	12.27
	CogCommon	**13.85**	**20.93**	7.02	10.39	**12.07**	**19.18**	**10.98**	**16.83**
llama3.1-8b-instruct	GraphRAG	9.09	13.16	6.98	9.52	5.49	8.26	7.19	10.31
	CogCommon	**11.32**	**16.22**	6.90	8.16	**7.14**	**10.00**	**8.45**	**11.46**
qwen2.5-7b-instruct	GraphRAG	8.33	13.68	12.07	17.95	8.70	14.29	9.70	15.31
	CogCommon	8.33	13.33	**13.04**	**20.22**	**11.58**	**17.60**	**10.98**	**17.05**
llama3.2-3b-instruct	GraphRAG	8.51	11.76	10.42	14.71	1.85	2.38	6.93	9.62
	CogCommon	**18.52**	**20.83**	10.34	12.24	**8.00**	**10.00**	**12.29**	**14.36**

3. Experiments

Our work aims to investigate the impact of cross-domain knowledge on entity and relationship extraction in knowledge graph construction, evaluating improvements in graph quality. We compare the performance of the proposed method with the GraphRAG [4]. Details are provided in Appendix B.

Datasets To construct a domain-specific knowledge graph, we crawl 15 papers for 3 domains, i.e., flatoptics, mmwave, and siphotonics from online publication venues such as IEEE Xplore. The original texts are converted to a structured format and pre-processed to remove charts, references, and formatting elements while preserving chapter structure and key terms. Additionally, we conduct supplementary experiments on Legalbench-RAG dataset [5] using RAG as the downstream task. Full results are provided in Appendix D.

Evaluation Indicators We evaluate the performance through two aspects: 1) **Content Quality**: Quantified via entity extraction accuracy and macro-F1 scores, providing robust metrics for information retrieval precision; 2) **Structural Quality**: Assessed through cross-domain ratio and modularity analysis [6], capturing topological characteristics of the resulting network. Detailed methodological specifications and metric calculations are elaborated in Appendix C.

Results and Analysis

Table 1 presents content quality metrics across three specialized domains (FlatOptics, mmWave, and SiPhotonics) from our semiconductor paper dataset. Figure 2 illustrates the structural metrics of the constructed knowledge graphs, specifically cross-domain connection ratio and modularity. Our findings are as follows:

Commonality knowledge benefits graph content quality. Our method demonstrates consistent improvements in entity extraction accuracy and Macro_F1 score across both large and small-scale LLMs. The observed results align with Scaling Laws[7], where larger models inherently perform better due to their greater capacity. However, we find that smaller models like Llama3.2-3b-instruct exhibit a higher relative improvement, suggesting that explicit cross-domain information can help bridge the capability gap between model sizes.

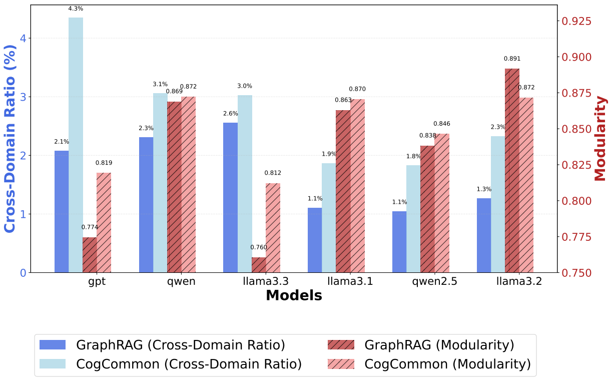

Fig. 2: Comparisons of structural quality of different models.

Cross-domain connections bring structural optimization. More accurate entity recognition directly enhances cross-domain connections. Llama3.3-70b-instruct demonstrates this relationship clearly: its 17.52% accuracy improvement corresponds with an 18.36% increase in cross-domain ratio. Simultaneously, our method slightly improves graph modularity, showing that CogCommon enhances domain connections while preserving community structure integrity.

4. Conclusion

Our approach enhances cross-domain connectivity by identifying shared cognitive structures, with smaller models benefiting significantly. By preserving structural modularity while strengthening domain links, CogCommon provides an efficient solution to knowledge fragmentation, with broader applications beyond semiconductors.

Acknowledgments

This research is supported by A*STAR Career Development Fund <Project No. C243512010>.

References

[1] Lutz Bornmann and Ruediger Mutz. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references, 2014.

[2] Sabine Hoffmann et al. Methods and procedures of transdisciplinary knowledge integration: Empirical insights from four thematic synthesis processes. *Ecology and Society*, 22(1), 2017. Accessed 10 Mar. 2025.

[3] George Lakoff and Mark Johnson. The metaphorical structure of the human conceptual system. *Cognitive Science*, 4(2):195–208, 1980.

[4] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization, 2024.

[5] Nicholas Pipitone and Ghita Houir Alami. Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain. *CoRR*, abs/2408.10343, 2024.

[6] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, June 2002.

[7] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.

[8] Hanzhu Chen, Xu Shen, Qitan Lv, Jie Wang, Xiaoqi Ni, and Jieping Ye. SAC-KG: exploiting large language models as skilled automatic constructors for domain knowledge graphs. *CoRR*, abs/2410.02811, 2024.

[9] Pranoy Panda, Ankush Agarwal, Chaitanya Devaguptapu, Manohar Kaul, and Prathosh A P. HOLMES: hyper-relational knowledge graphs for multi-hop question answering using llms. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 13263–13282. Association for Computational Linguistics, 2024.

[10] Xintao Wang, Qianwen Yang, Yongting Qiu, Jiaqing Liang, Qianyu He, Zhouhong Gu, Yanghua Xiao, and Wei Wang. Knowledgpt: Enhancing large language models with retrieval and storage access on knowledge bases. *CoRR*, abs/2308.11761, 2023.

[11] Jérôme Euzenat and Pavel Shvaiko. Ontology matching. In *Springer Berlin Heidelberg*, 2007.

[12] Zequn Sun, Muhao Chen, Wei Hu, Chengming Wang, Jian Dai, and Wei Zhang. Knowledge association with hyperbolic knowledge graph embeddings. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5704–5716, Online, November 2020. Association for Computational Linguistics.

[13] Ekaterina Shutova and Lin Sun. Unsupervised metaphor identification using hierarchical graph factorization clustering. In Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff, editors, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 978–988, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[14] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China, November 2019. Association for Computational Linguistics.

[15] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science, 2022.

[16] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599, 2024.

[17] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020.

[18] Qwen Team. Qwen2.5 technical report, 2025.

[19] Aaron Grattafiori, Abhimanyu Dubey, and Abhinav Jauhri et al. The llama 3 herd of models, 2024.

Appendix A. Related Work

LLMs have revolutionized knowledge graph construction with improved scalability and adaptability [8, 9, 10], yet typically focus on surface-level extraction within single domains. Traditional cross-domain integration approaches such as ontology alignment [11] and embedding methods [12] require extensive manual mapping or training data — particularly challenging in specialized domains with varied terminology. While cognitive linguistics has established that human understanding relies on systematic cross-domain mappings [3, 13], these insights remain largely unexploited in computational knowledge integration. Despite LLMs' strong knowledge capabilities [14], they struggle with specialized technical concepts [15], and recent LLM-KG integration work [16] primarily enhances LLMs rather than improving KG construction.

Appendix B. Baseline and Models

Baseline Implementation. We build our modifications on **nano-GraphRAG**[1].This project provides a smaller, faster implementation of GraphRAG. To ensure consistency, the chunk size is fixed at 1,000 across all datasets. Additionally, the gleaning parameter for GraphRAG is set to 3. All LLM implementations are based on the Transformers library [17], using a consistent set of hyperparameters for generation across all models.

Models. To comprehensively evaluate the effectiveness of our approach across different model scales and types, we selected both open source and closed source models. For closed source models, we included GPT-4o-mini[2] and Qwen-plus [3]. For open source models, we chose models with parameter scales ranging from 3b to 70b, including Qwen2.5 [18], and Llama3.1 [19],Llama3.2 [4],Llama3.3 [5].

Appendix C. Metrics

To evaluate the quality of entity and relation extraction in our knowledge graph framework, we employ two primary metrics: Accuracy and Macro F1 score.

3.0.1 Accuracy and Macro_F1

In the context of entity extraction from knowledge graphs, accuracy measures the proportion of correctly identified entities among all predicted entities, the F1 score provides a balanced measure of both the correctness and completeness of extracted entities.

Formally, in our result, accuracy is defined as:

$$\text{Accuracy} = \frac{TP_d}{TP_d + FP_d}$$

Macro_F1 is defined as:

$$\text{Macro F1} = \frac{2}{|D|} \sum_{d \in D} \frac{\text{Precision}_d \times \text{Recall}_d}{\text{Precision}_d + \text{Recall}_d}$$

where D is the set of all domains, and Precision_d and Recall_d are defined for each domain d as:

$$\text{Precision}_d = \frac{TP_d}{TP_d + FP_d}$$

$$\text{Recall}_d = \frac{TP_d}{TP_d + FN_d}$$

TP_d, FP_d, and FN_d represent true positives, false positives, and false negatives for domain d, respectively.

3.0.2 Structural Quality Metrics

To evaluate the topological properties of the constructed knowledge graph, particularly the quality of cross-domain connections and community structures, we utilize modularity and cross-domain ratio metrics, which are defined as follows:

[1] https://github.com/gusye1234/nano-graphrag

[2] https://openai.com/index/gpt-4o-system-card/

[3] https://www.alibabacloud.com/help/en/model-studio/developer-reference/what-is-qwen-llm

[4] https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/

[5] https://ai.meta.com/blog/meta-llama-3/

Modularity quantifies the strength of division of a network into communities. Higher modularity values indicate more distinct community structures with dense connections within communities and sparse connections between communities. For a graph $G = (V, E)$, modularity is defined as:

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

where A_{ij} represents the edge weight between nodes i and j, k_i and k_j are the degrees of nodes i and j respectively. m is the total number of edges in the network, c_i is the community to which node i is assigned, $\delta(c_i, c_j)$ is the Kronecker delta function, which equals 1 if $c_i = c_j$ and 0 otherwise.

In our implementation, we utilize the Louvain method for community detection, which iteratively optimizes the modularity value to identify community structures within the knowledge graph.

The cross-domain ratio measures the proportion of communities that span multiple domains, indicating the degree of integration between different knowledge domains. It is calculated as:

$$\text{CDR} = \frac{|\{C \in \mathcal{C} : |\{d(v) : v \in C\}| > 1\}|}{|\mathcal{C}|}$$

where $\mathcal{C}$ is the set of all detected communities, $d(v)$ represents the domain of node v, $|\{d(v) : v \in C\}|$ counts the number of distinct domains present in community C.

A higher cross-domain ratio indicates stronger interconnections between domains, suggesting effective knowledge integration across domain boundaries.

Appendix D. Full experiment results

Table A1: Comparisons of model performance on cross-domain Ratio, modularity, and RAG Task Metrics on LegalBench-RAG.

Model	Method	CDR(%)	Modularity	retrieval	
				Accuracy(%)	F1 (%)
gpt-4o-mini	GraphRAG	13.64	0.7423	0.1845	0.4655
	CogCommon	**22.22**	0.7329	**0.2424**	**0.4688**

The Table A1 presents performance of basic GraphRAG and our method across different models on LegalBench-RAG. Since the bench does not provide the ground truth extraction standard, the related metrics of entity relation extraction are not available. We compare retrieval effectiveness (accuracy, macro_F1) with knowledge graph structural properties (cross-domain Ratio, modularity). This allows for direct observation of how different approaches balance information retrieval performance with cross-domain knowledge integration capabilities.

Appendix E. Prompt Templates

Entity Extraction

Goal: Given a text document and a list of entity types, identify all entities and extract the following information for each identified entity:

- **name**: Name of the entity, capitalized.
- **type**: One of the following types: [entity_types].
- **description**: Comprehensive description of the entity's attributes and activities.

Format each entity as follows:

$$(”entity”\langle name\rangle\langle type\rangle\langle description\rangle)$$

Relation Extraction

Goal: Identify all pairs of (source_entity, target_entity) that are clearly related to each other. For each related entity pair:

- **source**: Name of the source entity.
- **target**: Name of the target entity.
- **description**: Explanation as to why you think the source entity and the target entity are related to each other.
- **strength**: Numeric score indicating the strength of the relationship.

Format each relationship as:

$$(”relationship”\langle source\rangle\langle target\rangle\langle description\rangle\langle strength\rangle)$$

Domain Commonality Discovery

Goal: Identify deep semantic commonalities across multiple domains.

I have analysis results from multiple different domains. Go beyond surface-level similarities and identify deep conceptual metaphors and high-level semantic patterns that connect these domains.

Domain analysis results:

```
{all_data}
```

Please identify:

1. **Conceptual Metaphors**: Identify how concepts from one domain are metaphorically mapped to another domain. For example, how ”argument” might be conceptualized as ”war” across domains.
2. **Structural Patterns**: Identify common structural relationships and patterns that appear across domains, even when the specific entities are different.
3. **Abstraction Hierarchies**: Identify how specific concepts across domains can be abstracted to more general patterns.
4. **Cognitive Frames**: Identify shared cognitive frames that organize experience across different domains.

Focus on DEEP patterns that would not be obvious from simple word matching. Look for how humans conceptualize these different domains using similar cognitive structures.

Respond in JSON format as follows:

```
{
    "conceptual_metaphors": [
        {
            "name": "metaphor name",
            "description": "detailed description",
            "source_domain": "domain providing the metaphorical structure",
            "target_domains": ["domain1", "domain2"],
            "examples": ["example1", "example2"]
        }
    ],
    "structural_patterns": [
        {
            "pattern": "pattern name",
            "description": "detailed description",
            "manifestations": ["manifestation in domain1", "manifestation in domain2"]
        }
    ],
    "abstraction_hierarchies": [
        {
            "abstract_concept": "high-level concept",
            "specific_manifestations": ["manifestation in domain1", "manifestation in domain2"]
        }
    ],
    "cognitive_frames": [
        {
            "frame": "frame name",
            "description": "detailed description",
            "elements": ["element1", "element2"],
            "domains": ["domain1", "domain2"]
        }
    ]
}
```