

RAG vs. GraphRAG: A Systematic Evaluation and Key Insights

📅 Announcement Date	@2025년 4월 17일
☰ Conference Name	arXiv 2025.02
⋮ Keywords	LLM RAG

1. Introduction

- **RAG의 활용:** RAG는 의료, 법률, 금융, 교육 등 다양한 분야에서 LLM의 사실성 (factuality), 안정성, 개인정보 보호 측면을 개선하는 데 기여해 왔습니다.
- **GraphRAG의 등장:** 기존 그래프 구조를 활용하거나 텍스트에서 그래프를 구성하여 복잡한 관계 및 논리를 표현할 수 있다는 점에서 주목받고 있습니다.
- **문제 제기:** 대부분의 GraphRAG 연구는 제한된 데이터셋과 작업에만 초점을 두며, **RAG와 GraphRAG를 동일 조건 하에서 비교한 연구는 거의 없습니다.**

Main Contributions

1. **체계적 평가:** 다양한 텍스트 기반 작업에서 RAG와 GraphRAG를 동일한 기준으로 비교.
2. **작업별 분석:** 각각의 방식이 특정 쿼리 유형에 대해 가지는 장점 도출.
3. **하이브리드 전략 제안:**
 - **선택적 활용(Selection):** 쿼리 성격에 따라 RAG 또는 GraphRAG 중 적합한 것을 선택.
 - **통합 활용(Integration):** 두 방식의 결과를 결합하여 최종 응답 생성.
4. **한계와 향후 방향:** GraphRAG의 구조적 제약과 이를 보완하기 위한 연구 방향 제시.

2. Related Work

2.1 RAG

- RAG는 LLM의 한계를 보완하기 위해 외부에서 정보를 검색하여 생성 결과를 향상시키는 기술이다.
- 다양한 검색 및 후처리 기법이 결합되어 QA, 요약, 대화 등에서 효과가 입증되었다.
- 그러나 RAG와 GraphRAG를 같은 조건에서 비교한 포괄적인 연구는 지금까지 없었다.

2.2 GraphRAG

- GraphRAG은 그래프 데이터 구조의 관계성을 활용하여 정보를 검색하는 방식이다.
- 기존에는 지식 그래프 기반 질의 응답이나 사실 검증 중심으로 사용되었고, 최근에는 텍스트로부터 그래프를 생성하여 활용하는 시도도 있다.
- 하지만 대부분 특정 작업에 한정되어 있으며, 일반적인 텍스트 작업에서의 효과와 적용 조건은 명확히 정리되지 않았다.

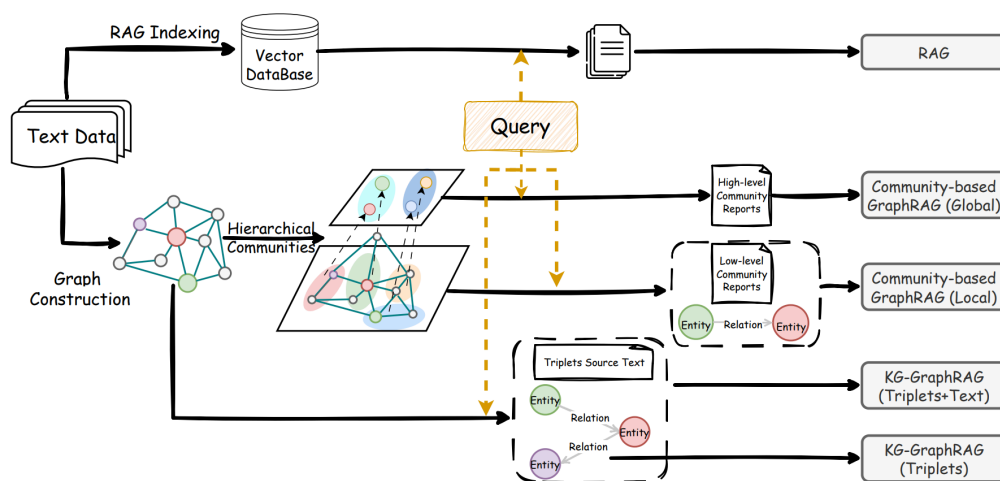


Figure 1: The illustration of RAG, KG-based GraphRAGs and Community-based GraphRAGs.

RAG vs. GraphRAG 구조 흐름도

📄 ① 입력: Text Data (텍스트 데이터)

- 논문이나 문서와 같은 일반적인 텍스트를 의미합니다.
- 이 데이터는 아래 두 경로로 분기됩니다.

② RAG 경로: Vector DB 기반

- 텍스트를 청크로 분할 → 임베딩 → 벡터 DB에 저장
- 사용자가 쿼리를 입력하면, 벡터 유사도 검색을 통해 관련 청크를 찾아 RAG가 응답 생성

👉 전통적 의미 기반 RAG 방식

③ GraphRAG 경로: 그래프 구성

- 텍스트 데이터를 LLM 등을 통해 그래프로 변환
- 엔티티(Entity)와 관계(Relation)를 추출하여 그래프 노드/엣지로 구성

이후 두 갈래로 다시 분기합니다:

◆ KG 기반 GraphRAG (하단 오른쪽)

- (1) **Triplets only**: 단순한 삼항 구조 (예: [A] → [관계] → [B]) 만 사용
- (2) **Triplets + Text**: 위 트리플렛과 함께 원본 텍스트도 함께 검색

➡ 이 두 방식은 그래프 구조만으로 정보를 추출하는 방식입니다.

◆ 커뮤니티 기반 GraphRAG (중간 오른쪽)

- 그래프에서 ****계층적 커뮤니티(Hierarchical Communities)****를 감지하고 구조화
- 각 커뮤니티에 대해 요약 리포트를 생성







두 가지 검색 방식 제공:

1. **Global**: 상위 커뮤니티 요약만 검색 (포괄적, 요약형)
2. **Local**: 하위 커뮤니티 + 엔티티 + 관계까지 세부적으로 검색

➡ **Community-GraphRAG(Global)** 과 **(Local)** 이 각각 해당됩니다.

Query 흐름

- 사용자의 질의(Query)는 위의 각 구성 방식에 따라 각각 검색 및 응답 생성에 사용된다.
- 점선 화살표가 Query 경로를 나타냅니다.

항목	Knowledge Graph 기반 (KG-GraphRAG)	Community-based GraphRAG
 그래프 구성 방식	LLM을 이용해 삼항 구조 (triplets) 추출 → (주어, 관계, 목적어)	LLM으로 엔티티 및 관계 추출 후 그래프 커뮤니티 탐지 알고리즘으로 계층 구조 형성
 검색 방식	질의의 엔티티를 기반으로 그래프 내에서 탐색 (예: multi-hop으로 연결된 이웃 노드 추적)	두 가지 검색 방식 존재: ① Local Search : 질의와 일치하는 엔티티, 관계, 하위 커뮤니티 ② Global Search : 의미 유사도 기반 상위 커뮤니티 요약 검색
 추출 정보	- Triplet만 추출하거나- Triplet + 관련 원문 텍스트까지(2가지 버전 존재)	- 계층적 커뮤니티 리포트 (요약본)- 엔티티/관계/요약을 함께 결합 가능
 장점	- 추론 가능한 구조- 관계 기반 탐색이 명확함	- 다양한 레벨에서 요약 제공 가능- 정보의 다양성/종합성 확보에 유리
 한계	- 그래프 불완전성(누락된 엔티티/관계)- 정교한 KG 생성이 어려움	- Global search는 세부 정보 부족- 요약 순서에 따른 LLM 편향 문제 발생
 적합한 작업	- 정형적 추론이 필요한 질의 응답 (예: 엔티티 연결 중심 QA)- 지식 검증	- 다층적 요약, 복잡한 질의- 멀티홉 QA , 쿼리 중심 요약, 문맥 비교

3. Evaluation Methodology

3.1 RAG

- RAG 시스템은 텍스트를 256 토큰 단위로 청크하고 의미 유사도 기반 검색을 수행합니다.
- OpenAI의 text-embedding-ada-002 모델로 청크를 임베딩하고 Top-10 유사 청크를 검색합니다.
- 생성기는 Llama-3.1-8B 및 70B 모델을 사용하여 결과 응답을 생성합니다.
- 단일 문서와 다중 문서 작업에 따라 개별 또는 통합 RAG 인덱스를 사용합니다.

3.2 GraphRAG

- 두 가지 GraphRAG 방식인 KG 기반 GraphRAG과 커뮤니티 기반 GraphRAG이 평가됩니다.

- KG-GraphRAG은 텍스트로부터 엔티티와 관계를 추출해 지식 그래프를 만들고 이를 통해 정보를 검색합니다.
- Community-GraphRAG은 계층적 커뮤니티를 구성하고, Local(엔티티 중심)과 Global(요약 중심) 검색 방식을 제공합니다.
- 실험의 일관성을 위해 RAG와 동일한 임베딩 모델, 청크 전략, LLM을 적용하여 비교합니다.

4. Question Answering

질문 응답(QA)은 RAG 시스템 성능 평가에서 가장 널리 사용되는 작업 중 하나입니다. 이 장에서는 **단일 홉 / 다중 홉 QA**, **단일 문서 / 다중 문서 시나리오**에서 RAG와 GraphRAG의 성능을 평가합니다. 총 네 가지 대표적인 데이터셋을 사용하며, 세 가지 주요 측정 지표(정확도, 정밀도, 재현율, F1)를 적용합니다.

4.1 Datasets and Evaluation Metrics

데이터셋

1. **NQ (Natural Questions)**: Google Search의 실제 질문을 기반으로 한 단일 홉 QA 데이터셋
2. **HotpotQA**: 다중 문서 기반 멀티홉 QA, 난이도 높은 브릿지 질문 중심
3. **MultiHop-RAG**: 뉴스 기사 기반 멀티도큐먼트 QA. 질의 유형: 추론, 비교, 시간, Null
4. **NovelQA**: 소설 기반의 긴 텍스트 QA. 다양한 질문 유형(21종)을 포함하며 고난도

평가지표

- NQ/Hotpot: **정밀도(P), 재현율(R), F1 Score**
- MultiHop-RAG/NovelQA: **정확도 (Accuracy)**

4.2 QA Main Results

Table 1: Performance comparison (%) on NQ and Hotpot datasets. The best results are highlighted in bold, and the second-best results are underlined.

Method	NQ						Hotpot					
	Llama 3.1-8B			Llama 3.1-70B			Llama 3.1-8B			Llama 3.1-70B		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
RAG	71.7	63.93	64.78	74.55	67.82	68.18	62.32	60.47	60.04	66.34	63.99	63.88
KG-GraphRAG (Triplets only)	40.09	33.56	34.28	37.84	31.22	28.50	26.88	24.81	25.02	32.59	30.63	30.73
KG-GraphRAG (Triplets+Text)	58.36	48.93	50.27	60.91	52.75	53.88	45.22	42.85	42.60	51.44	48.99	48.75
Community-GraphRAG (Local)	<u>69.48</u>	<u>62.54</u>	<u>63.01</u>	<u>71.27</u>	<u>65.46</u>	<u>65.44</u>	64.14	62.08	61.66	67.20	64.89	64.60
Community-GraphRAG (Global)	60.76	54.99	54.48	61.15	55.52	55.05	45.72	47.60	45.16	48.33	48.56	46.99

Table 2: Performance comparison (%) on the MultiHop-RAG dataset across different query types.

Method	Llama 3.1-8B					Llama 3.1-70B				
	Inference	Comparison	Null	Temporal	Overall	Inference	Comparison	Null	Temporal	Overall
RAG	92.16	57.59	96.01	30.7	<u>67.02</u>	94.85	56.31	91.36	25.73	<u>65.77</u>
KG-GraphRAG (Triplets only)	55.76	22.55	98.67	18.7	41.24	76.96	32.36	94.35	19.55	50.98
KG-GraphRAG (Triplets+Text)	67.4	34.7	97.34	17.15	48.51	85.91	35.98	86.38	21.61	54.58
Community-GraphRAG (Local)	86.89	<u>60.63</u>	80.07	50.6	69.01	92.03	<u>60.16</u>	<u>88.70</u>	<u>49.06</u>	71.17
Community-GraphRAG (Global)	<u>89.34</u>	64.02	19.27	53.34	64.4	<u>89.09</u>	66.00	13.95	59.18	65.69

Table 3: Performance comparison (%) on the NovelQA dataset across different query types with LLama 3.1-8B.

RAG									KG-GraphRAG (Triplets+Text)								
	chara	mean	plot	relat	settg	span	times	avg		chara	mean	plot	relat	settg	span	times	avg
mh	68.75	52.94	58.33	75.28	92.31	64.00	33.96	47.34	mh	52.08	52.94	44.44	55.06	69.23	64.00	28.61	38.37
sh	69.08	62.86	66.11	75.00	78.35	-	-	68.73	sh	36.84	45.71	40.17	87.50	36.08	-	-	39.93
dtl	64.29	45.51	78.57	10.71	83.78	-	-	55.28	dtl	38.57	30.90	42.86	21.43	32.43	-	-	33.60
avg	67.78	50.57	67.37	60.80	80.95	64.00	33.96	57.12	avg	40.00	36.23	41.09	49.60	38.10	64.00	28.61	37.80
Community-GraphRAG (Local)									Community-GraphRAG (Global)								
	chara	mean	plot	relat	settg	span	times	avg		chara	mean	plot	relat	settg	span	times	avg
mh	68.75	64.71	55.56	67.42	92.31	52.00	35.83	47.01	mh	54.17	58.82	55.56	56.18	53.85	68.00	20.59	34.39
sh	59.87	58.57	65.69	87.50	64.95	-	-	63.43	sh	45.39	50.00	55.65	87.50	38.14	-	-	49.65
dtl	54.29	37.64	62.50	25.00	70.27	-	-	46.88	dtl	28.57	29.78	32.14	87.50	40.54	-	-	30.89
avg	60.00	44.91	64.05	59.20	68.71	52.00	35.83	53.03	avg	42.59	36.98	51.66	52.00	40.14	68.00	20.59	39.17

데이터셋	LLM 모델	최고 성능 방식
NQ (단일홉)	Llama 3.1-70B	RAG
HotpotQA (멀티홉)	Llama 3.1-70B	Community-GraphRAG (Local)
MultiHop-RAG	Llama 3.1-70B	Community-GraphRAG (Local), 일부 Temporal에선 Global
NovelQA	Llama 3.1-8B	혼합: RAG(단일홉/세부정보), GraphRAG(멀티홉)

주요 관찰:

1. **RAG**는 단일 홉 및 세부 정보 중심 질의에 강함.
2. *Community-GraphRAG (Local)*은 멀티홉, 추론 중심 질문에서 우수.
3. **Global** 방식은 요약 위주로 정확도는 낮고, 환각(hallucination) 위험이 있음.
4. **KG 기반 방식은 전반적으로 낮은 성능** – 이유는 그래프에 포함되지 않은 답변 엔티티가 많기 때문 (NQ 65.5%, Hotpot 65.8%)

4.3 Comparative QA Analysis

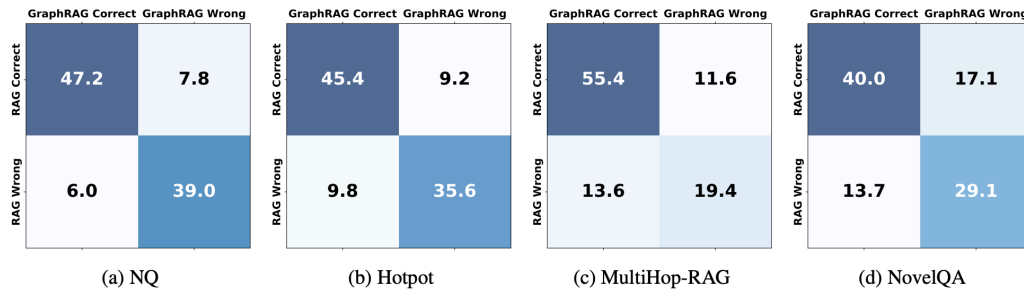


Figure 2: Confusion matrices comparing GraphRAG and RAG correctness across datasets using Llama 3.1-8B.

이 절에서는 **정답 여부에 따른 질의 분류**를 통해 각 방식의 상대적인 강점을 정량적으로 분석합니다.

- 카테고리:
 1. 둘 다 정답
 2. RAG만 정답 (RAG-only)
 3. GraphRAG만 정답 (GraphRAG-only)
 4. 둘 다 오답

 예: MultiHop-RAG 데이터셋 기준

- RAG-only: 11.6%, GraphRAG-only: 13.6%
 - 각 방식이 서로 다른 질의 유형에 강점이 있음을 시사

결론: 혼합 활용 전략의 필요성 강조

4.4 Improving QA Performance

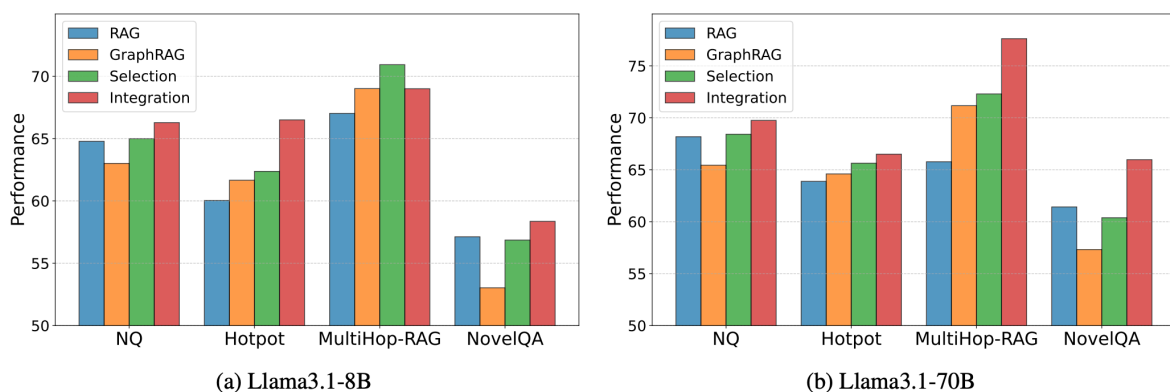



Figure 3: Overall QA performance comparison of different methods.

 제안된 두 가지 전략:

1. 선택적 실행 전략 (Selection Strategy)

- 질의를 분류해 **Fact-based → RAG, Reasoning-based → GraphRAG(Local)**
- LLM의 in-context learning을 활용해 쿼리 유형 자동 분류
- 장점: 효율적, 단점: 성능은 통합보단 낮음

2. 통합 실행 전략 (Integration Strategy)

- RAG와 GraphRAG 모두에서 정보를 검색해 **결과를 결합하여 생성**
- 성능이 가장 높지만, 계산 자원이 많이 듦

 결과: MultiHop-RAG에서 최대 6.4% 성능 향상

→ 통합 방식이 최상 성능, 선택 방식은 효율성 우위

5. Query-Based Summarization

Query-based 요약은 RAG 를 평가하는데 널리 사용되는 task임.

5.1 Datasets and Evaluation Metrics

- 본 절에서는 쿼리 기반 요약 평가를 위해 네 가지 대표적인 데이터셋을 사용
- 단일 문서용으로는 SQuALITY와 QMSum, 다중 문서용으로는 ODSum-story 및 ODSum-meeting이 활용됨
- 각 쿼리는 특정 역할, 사건, 인물 등에 초점을 둔 질문 형식임
- 평가 지표로는 표면적 일치율(ROUGE-2)과 의미 유사도(BERTScore)를 사용

5.2 Summarization Experimental Results

Table 4: The performance of query-based single document summarization task using Llama3.1-8B.

Method	SQuALITY						QMSum					
	ROUGE-2			BERTScore			ROUGE-2			BERTScore		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
RAG	15.09	8.74	10.08	74.54	81.00	77.62	<u>21.50</u>	3.80	<u>6.32</u>	81.03	<u>84.45</u>	82.69
KG-GraphRAG (Triplets only)	11.99	6.16	7.41	82.46	84.30	83.17	13.71	2.55	4.15	80.16	82.96	81.52
KG-GraphRAG (Triplets+Text)	15.00	9.48	<u>10.52</u>	84.37	85.88	84.92	16.83	3.32	5.38	<u>80.92</u>	83.64	82.25
Community-GraphRAG (Local)	15.82	8.64	10.10	<u>83.93</u>	<u>85.84</u>	<u>84.66</u>	20.54	3.35	5.64	80.63	84.13	82.34
Community-GraphRAG (Global)	10.23	6.21	6.99	82.68	84.26	83.30	10.54	1.97	3.23	79.79	82.47	81.10
Integration	<u>15.69</u>	<u>9.32</u>	10.67	74.56	81.22	77.73	21.97	3.80	6.34	80.89	84.47	<u>82.63</u>

Table 5: The performance of query-based multiple document summarization task using Llama3.1-8B.

Method	ODSum-story						ODSum-meeting					
	ROUGE-2			BERTScore			ROUGE-2			BERTScore		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
RAG	15.39	<u>8.44</u>	9.81	83.87	85.74	84.57	15.50	6.43	8.77	83.12	85.84	84.45
KG-GraphRAG (Triplets only)	11.02	5.56	6.62	82.09	83.91	82.77	11.64	4.87	6.58	81.13	84.32	82.69
KG-GraphRAG (Triplets+Text)	9.19	5.82	6.22	79.39	83.30	81.03	11.97	4.97	6.72	81.50	84.41	82.92
Community-GraphRAG (Local)	<u>13.84</u>	7.19	8.49	83.19	85.07	83.90	<u>15.65</u>	5.66	8.02	82.44	85.54	83.96
Community-GraphRAG (Global)	9.40	4.47	5.46	81.46	83.54	82.30	11.44	3.89	5.59	81.20	84.50	82.81
Integration	14.77	8.55	<u>9.53</u>	<u>83.73</u>	<u>85.56</u>	<u>84.40</u>	15.69	<u>6.15</u>	<u>8.51</u>	<u>82.87</u>	<u>85.81</u>	<u>84.31</u>

단일 문서 요약 결과 (Llama 3.1-8B 기준)

- RAG가 대부분 좋은 성능을 기록
- KG-GraphRAG은 Triplet만 쓸 때보다, **Triplet+Text**가 성능이 높음
- Community-GraphRAG(Local)은 RAG와 비슷하거나 약간 낮은 수준

다중 문서 요약 결과

- RAG와 Community-GraphRAG(Local)가 가장 우수
- Global 방식은 모든 경우에서 낮은 성능
- Integration 전략은 대체로 **RAG 단독 성능에 근접하거나 약간 향상**

5.3 Position Bias in Evaluation

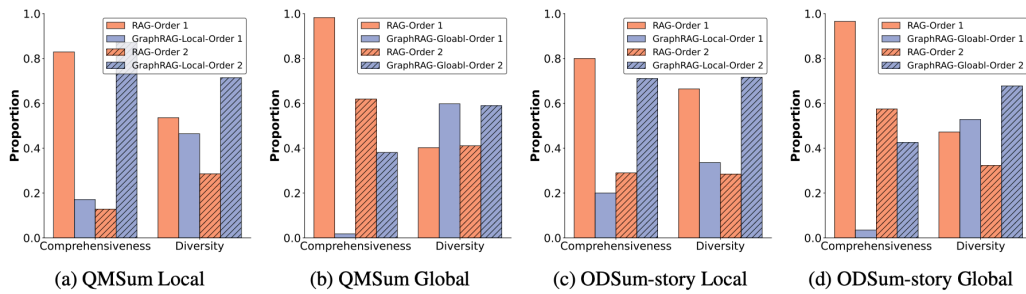


Figure 4: Comparison of LLM-as-a-Judge evaluations for RAG and GraphRAG. "Local" refers to the evaluation of RAG vs. GraphRAG-Local, while "Global" refers to RAG vs. GraphRAG-Global.

Edge et al. (2024)는 LLM에게 두 개의 요약 결과를 보여주고 어느 쪽이 더 좋은지 선택하게 하는 방식(LLM-as-a-Judge)을 사용했는데, 이 방식은 요약이 먼저 등장하느냐 나중에 등장하느냐에 따라 결과가 달라지는 편향(position bias)을 초래함

실험 세팅:

- 요약 A가 앞에, 요약 B가 뒤에 나오는 순서를 바꾸며 2회 평가
- 평가 기준: **Comprehensiveness**(세부정보 충실도), **Diversity**(다양성/전체 시각)

결과:

- RAG vs GraphRAG(Local): 위치 순서 바꾸면 정반대 결과 나옴
- RAG vs GraphRAG(Global): Comprehensiveness는 RAG 우세, Diversity는 GraphRAG(Global) 우세

결론: LLM-as-a-Judge는 신뢰도에 한계, 정량적 지표(BERTScore 등) 사용 필요

6. Conclusion & Limitations

6.1 Conclusion

- 본 연구는 RAG와 GraphRAG를 동일한 조건에서 평가한 최초의 체계적 비교 분석입니다.
- 두 방식은 각각 강점이 다르며, 작업 유형에 따라 효과가 달라집니다.
- RAG는 세부 정보 검색에 강하고, GraphRAG(Local)는 멀티홉 추론에 강했습니다.
- Global 방식은 다양성에는 기여했지만, 정보 정확도 측면에서 떨어졌습니다.
- 이 연구는 적절한 방식 선택 또는 통합 전략을 통해 RAG 시스템 성능을 향상시킬 수 있음을 보여줍니다.

6.2 Limitations

1. 그래프 구성 자동화의 어려움

- 현재 대부분의 GraphRAG 시스템은 수동으로 정의된 개체 추출 및 관계 모델에 의존하고 있으며, 이로 인해 그래프 누락과 부정확성이 발생합니다.

2. 질의 유형 분류 자동화 한계

- 하이브리드 전략 중 “선택 실행” 방식은 질의 유형을 정확히 분류해야 하는데, 이는 현 LLM의 in-context 분류 능력에 크게 의존하며 신뢰성이 낮을 수 있습니다.

3. 모든 작업에서 LLM-as-a-Judge는 사용되지 않음

- 일부 QA/요약 결과는 사람이 아닌 정량 지표 또는 사전 기준에 따라 평가되어 **사용자 선호 반영이 부족**할 수 있습니다.

4. 계산 비용

- 통합 전략은 성능이 가장 우수하지만, **LLM 2회 호출 + 이중 검색**으로 비용이 높아 실제 상용화에는 부담