KARMA: Multi-Agent LLMS for Automated Knowledge Graph Enrichment

다중 에이전트 LLM을 활용한 자동화된 지식 그래프 확장 프레임워크

Yuxing Lu(PKU, Georgia Tech) | Wei Wu(PKU) | Xukai Zhao(Tsinghua) | Rui Peng(PKU) | Jinzhuo Wang(PKU, †)

NeurIPS 2025

DeepShark Lab 학부연구생 김승겸

목차

지식 그래프(KG) 확장의 병목 현상

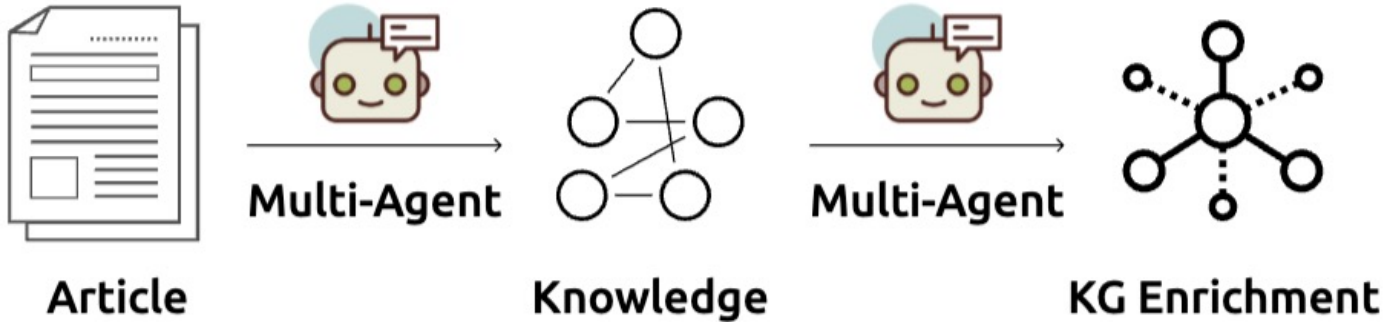
- **지식 그래프(KG)**: 다양한 분야의 복잡한 정보를 구조화 및 추론하는 데 필수적
- 매년 700만 개 이상 출판되는 과학 문헌
- **문제**

  : 기하급수적으로 증가하는 비정형 텍스트 지식을 구조화된 KG로 표현하는 데 격차 발생
- **기존 접근법의 한계**
  - **수동 큐레이션**: 정확도는 높지만, 대규모 데이터의 대규모 확장 불가
  - **기존 NLP 기술**: 특정 도메인별 용어 및 문맥 의존적 관계 처리에 취약함
- 기존 LLM 활용의 단점
  - **환각**: 복잡한 관계 추출 중 존재하지 않는 정보 생성
  - **일관성 부족**: 문서 간 스키마 일관성 유지 못 함
  - **비용 문제**: 전체 텍스트 처리 시 이차 계산 비용 발생

KARMA

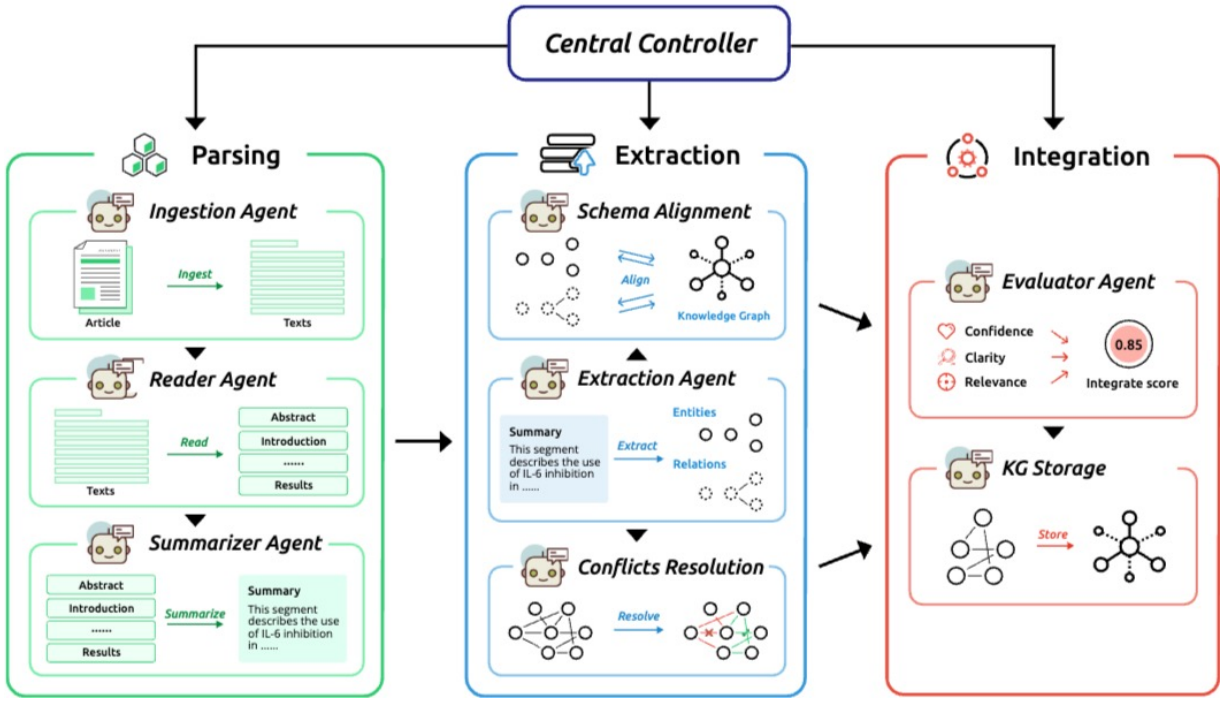
- **9개의 특화된 LLM 기반 에이전트**들이 협업하여 지식 추출 및 통합을 수행하는 **계층적 다중 에이전트 시스템**

- 복잡한 작업을 **모듈식 하위 작업**으로 분해하여 처리



KARMA

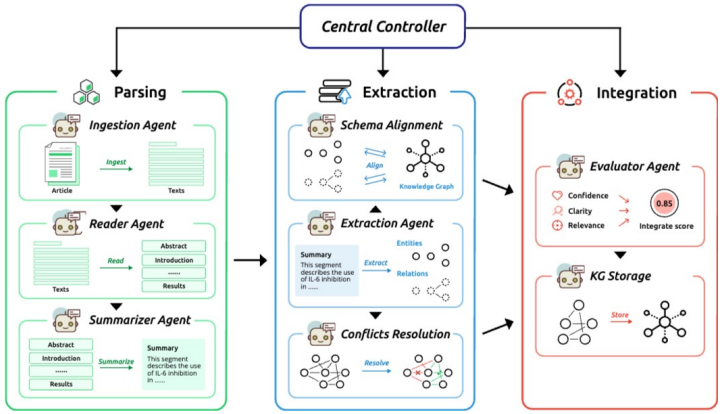
- **에이전트 간 상호 검증을 통해 추출**된 지식의 신뢰성 향상
- **도메인 적응형 프롬프트**: 정확도를 유지하면서 특화된 맥락 처리
- **모듈식 설계**: 새로운 엔터티 및 관계 등장 시 동적 업데이트를 지원하는 확장성 보장



KARMA – 파이프라인 3단계

1. 파싱(Parsing)

- **I**ngestion **A**gents: 입력 문서 검색 후 정규화
- **R**eader **A**gents: 관련 텍스트 섹션 파싱 후 분할
- **S**ummarizer **A**gents: 관련 섹션을 더 짧은 도메인별 요약으로 압축



2. 추출(Extraction)

- **E**ntity **E**xtraction **A**gents: 주제 관련 엔터티 식별 후 정규화
- **R**elationship **E**xtraction **A**gents: 엔터티 간 관계 추론
- **S**chema **A**lignment **A**gents: 개체와 관계를 KG 스키마에 정렬
- **C**onflict **R**esolution **A**gents: 기존 지식과의 논리적 불일치 감지 및 해결

3. 통합(Integration)

- **E**valuator **A**gents: 여러 검증 신호(신뢰도, 명확성, 관련성) 집계 후 최종 통합 여부 결정

KARMA – 문제 정의 (수식)

$\mathbf{G} = (\mathbf{V}, \mathbf{E})$ - 기존 지식 그래프

$\mathbf{t} = (\mathbf{e_h}, \mathbf{r}, \mathbf{e_t})$ - 관계 트리플렛

$\mathbf{G_{new}} = \mathbf{G} \cup (\bigcup_{i=1}^{n} \mathbf{K_i})$ - 확장된 지식 그래프

$\mathbf{K_i} = \mathbf{Extract}(\mathbf{p_i})$ - 문서 p_i에서 추출된 트리플렛 집합

V: 엔티티 집합 | E: 관계(에지) 집합 | e_h: 머리 엔티티 | e_t: 꼬리 엔티티 | r: 관계 유형

$$\mathbf{Extract}(\mathbf{p_i}) = \{\mathbf{t} \mid \mathbf{t} \notin \mathbf{E}, \mathbf{t} = (\mathbf{e_h}, \mathbf{r}, \mathbf{e_t}), \mathbf{e_h}, \mathbf{e_t} \in \mathbf{V}^{+}\}$$

$\forall \mathbf{t} \in \mathbf{K_i} : \mathbf{LLM_{verify}}(\mathbf{t}) = \mathbf{true}$ - LLM 기반 검증

$\mathbf{conflict}(\mathbf{t}, \mathbf{G}) = \mathbf{0}$ - 모순 해결

새로운 문서에서 유효한 트리플을 추출하고, LLM으로 검증 후 KG에 통합

KARMA – 문제 정의 (수식)

파싱

- $\mathrm{IA}(p_i) = (normalize(p_i), metadata(p_i))$

- p_i: 원시 출판물
- $normalize(p_i)$: LLM 프롬프트 Pingest 사용하여 OCR 오류 및 복잡성(구조적 불일치) 처리
- Output: 표준화된 텍스트 표현 및 주요 메타데이터(저널, 날짜, 저자 등)

- $R(s_j) = \mathrm{LLM}_{\mathrm{reader}}(s_j, \mathcal{G})$

- $R(s_j)$: 각 세그먼트의 관련성 점수
- $\mathrm{LLM}_{\mathrm{reader}}$: LLM 모델
- s_j: 분석 대상 세그먼트

- $u_j = \mathrm{LLM}_{\mathrm{summ}}(s_j, P_{\mathrm{summ}})$

- u_j: 계산 오버헤드를 줄이기 위해 s_j를 압축한 표현
- P_{summ}: LLM이 중요 개체, 관계 및 도메인별 용어를 유지하도록 하는 프롬프트

KARMA – 문제 정의 (수식)

추출

- $E(u_j) = \text{LLM}_E(u_j, P_E) \odot D_E$

- LLM_E: 프롬프트 P_E를 가진 전문화된 개체 추출 LLM
- $\odot D_E$: 사전/온톨로지 기반 필터링

- $\hat{e} = \text{argmin}_{v \in V} d(\phi(e), \psi(v))$

- $\phi(e), \psi(v)$: 텍스트 멘션 및 기존 KG 그래프 노드를 동일 공간으로 변환하는 임베딩 함수

- $p(r \mid \hat{e}_i, \hat{e}_j, u_j) = \text{LLM}_R(\hat{e}_i, \hat{e}_j, u_j, P_R)$

- $p(r \mid \cdot)$: 주어진 문맥 안에서 두 엔터티 간에 해당 관계가 성립할 확률

- $\tau^* = \arg\max_{\tau \in \mathcal{T}} \text{LLM}_{\text{SAA}}(v, \tau, P_{\text{align}})$

- v: 새로운 개체
- τ: 유효한 개체 유형(질병, 약물, 유전자 등)의 집합

- LLM_{SAA}: v가 τ에 속할 확률 추정

- $conflict(t, \mathcal{G}) = \begin{cases} 1, & if \exists t' \; that \; contradicts \; t, \\ 0, & otherwise \end{cases}$

- $\text{LLM}_{\text{CRA}}(t, t') \rightarrow \{\text{Agree}, \text{Contradict}\}$

KARMA – 문제 정의 (수식)

통합

- Confidence: $C(t) = \sigma(\sum \alpha_i v_i(t))$
- Clarity: $Cl(t) = \sigma\left(\sum \beta_j c_j(t)\right)$
- $Relevance$: $R(t) = \sigma(\sum \gamma_k r_k(t))$.
- integrate$(t) = \begin{cases} 1, & if \frac{C(t)+Cl(t)+R(t)}{3} \\ 0, & otherwise \end{cases}$

실험 및 평가 방법 – (1) 실험 데이터셋

세 가지 생물의학 도메인(PubMed 과학 출판물)

- 유전체학 (Genomics) 720편
- 단백질체학 (Proteomics) 360편
- 대사체학 (Metabolomics) 120편

- 모든 문서는 PDF 형식으로 저장, KARMA 내의 IA에 의해 처리됨

실험 및 평가 방법 – (2) 활용된 LLM 백본

- **GLM-4**: 90억 파라미터 오픈 소스 모델
- **GPT-4o**: RHLF(인간 피드백 기반 강화학습)을 통해 최적화된 독점적인 멀티모달 모델, 과학 지식 추출 및 개념 기반화에서의 강력한 적응성
- **DeepSeek-v3**: 370억 활성화 파라미터 MoE 오픈 소스 모델

실험 및 평가 방법 – (3) 평가 지표

1. 핵심 지표

- **평균 신뢰도**: 추출된 지식의 신뢰성
- **평균 명확성**: 관계가 얼마나 모호하지 않고 직접적인지
- **평균 관련성**: 새로 추가된 triplet이 KG의 도메인이나 범위와 얼마나 관련있는지

2. 그래프 통계

- **커버리지 증가량**: 이전 KG에 존재하지 않던 새로 도입된 엔터티의 수 측정(KG의 확장된 범위)
- **연결성 증가량**: 기존 엔터티들의 노드 차수(degree)의 총 증가량 계산(향상된 상호 연결성)

3. 품질 지표

- **충돌 비율**: 모순으로 인해 제거된 새로 추출된 엣지(관계)의 비율
- **LLM 기반 정확도**
- **질의응답 일관성**

주요 실험 결과 – 성능 평가 및 도메인 별 성능 비교

- GML-4 기반 단일 에이전트 방식 대비 우수한 성능
- 유전체학 도메인에서 최대 38,230개의 신규 엔터티 식별
- 대사체학 대비 3.6배 높은 커버리지 증가량 달성

주요 실험 결과 – LLM 백본별 성능 비교

- DeepSeek-v3: 전체 평가지표 24개 중 17개(71%)에서 우수한 성능 및 커버리지 극대화
- GPT-4o: 정밀도에 특화되어 가장 높은 LLM 기반 정확도(R_{LC}) 기록
- GLM-4: 대사체학에서 높은 정보 명확성(M_{Cla})에 강점

주요 실험 결과 – Ablation Study

- **Summarizer 에이전트 제거**: 노이즈 증가로 인해 정확도 대폭 하락
- **Conflict Resolution 에이전트 제거**: 논리적 모순을 걸러내지 못해 정확성 저하
- **Evaluator 에이전트 제거**: 낮은 신뢰도의 엣지가 포함되어 답변 품질이 저하

주요 실험 결과

Domain	Model	Core Metrics			Graph Stats.		Quality Indicators			
		$M_{\text{Con}}^{\uparrow}$	$M_{\text{Cla}}^{\uparrow}$	$M_{\text{Rel}}^{\uparrow}$	$\Delta_{\text{Cov}}^{\uparrow}$	$\Delta_{\text{Con}}^{\downarrow}$	$R_{\text{CR}}^{\uparrow}$	$R_{\text{LC}}^{\uparrow}$	$C_{\text{QA}}^{\uparrow}$	$R_{\text{HE}}^{\uparrow}$
Genomics	Single-Agent	NA	NA	NA	4384	1.083	NA	0.493	0.472	0.320
	GLM-4	0.729	0.804	**0.716**	4969	1.131	0.238	0.623	0.589	0.445
	GPT-4o	0.843	0.744	0.640	9795	1.265	**0.148**	**0.880**	0.569	0.510
	DeepSeek-v3	**0.846**	**0.754**	0.667	**38230**	**1.765**	0.186	0.831	**0.612**	**0.625**
Proteomics	Single-Agent	NA	NA	NA	5002	1.150	NA	0.638	0.572	0.415
	GLM-4	0.731	0.752	0.609	6832	1.173	0.214	0.720	**0.617**	0.500
	GPT-4o	0.823	0.797	0.613	7008	1.191	0.160	0.740	0.612	0.550
	DeepSeek-v3	**0.845**	**0.825**	**0.682**	**11936**	**1.468**	**0.151**	**0.772**	0.613	**0.575**
Metabolomics	Single-Agent	NA	NA	NA	485	1.077	NA	0.527	0.450	0.455
	GLM-4	0.701	**0.790**	0.762	703	1.159	0.188	0.617	0.449	0.485
	GPT-4o	**0.802**	0.730	0.726	773	1.143	0.147	**0.683**	0.482	0.535
	DeepSeek-v3	0.790	0.746	**0.767**	**1752**	**1.811**	**0.132**	0.668	**0.493**	**0.580**

연구 요약 및 의의

- KARMA는 과학 문헌으로부터 **지식 그래프 확장을 자동화**하고 **대규모 확장성을 제공**하는 **다중 에이전트 LLM 프레임워크**

- 작업을 **특화된 에이전트**(엔터티 발견, 관계 검증, 충돌 해결 등)로 **분할**하고 **교차 검증**을 수행하여 단일 에이전트 방식의 한계를 극복

연구의 한계 및 향후 과제

한계

- 전문가 검증의 부재
- 도메인별 성능 편차

향후 과제

- 하이브리드 시스템 도입
- 시스템 최적화

감사합니다