폐쇄망 환경을 위한 온톨로지 기반 로컬 LLM 시스템

이기호*, 이승주*, 이주찬**, 이용준**
*극동대학교 인공지능보안학과
**극동대학교 해킹보안학과
e-mail:kdu-kh@kdu.ac.kr*
dltmdwn0305@gmail.com*
jasonjennie3@gmail.com**
2020032@kdu.ac.kr**

An Ontology-Based Local LLM Framework for Air-Gapped Environments

Ki-Ho Lee*, Seung-Ju Lee*, Joo-Chan Lee**, Yong-Joon Lee**
*Dept. of Artificial Intelligence Security, Far East University
*Dept. of Artificial Intelligence Security, Far East University
**Dept. of Hacking Security, Far East University
**Dept. of Hacking Security, Far East University

요 약

본 논문에서는 외부 네트워크 연결이 제한된 폐쇄망 환경을 위해 온톨로지 로컬 RAG와 LLM 파이프라인을 설계 및 구현한다. 전처리 단계에서 PDF와 TXT 문서로부터 헤더와 푸터 제거, 문단과 문장 청킹, 키워드 추출을 수행하고, 약어 및 라벨 패턴을 자동 채굴한다. 검색 단계에서는 TF-IDF 인덱스 위에 온톨로지 기반의 질의 확장과 개념 적합도 우선 재랭킹을 적용해 컨텍스트 선택의 안정성을 높였으며 추론 단계는 llama 기반 GGUF 모델로 오프라인에서 응답을 생성한다. 평가는 스마트 그리드 환경과 관련한 문서를 기반으로 질문을 준비하였고, 로컬 GPT와 ChatGPT를 동일 질의 및 지침으로 비교하였다. 정량 지표에 더해 두 응답의 의미적 근접성을 인간 평가로 측정한 결과 전체 평균은 2.81점 수준으로 나타났다. 다만, 한국어 기반 질의에서는 유사도 4.13점으로 높았으며, 영어 기반 질의에서 다소 낮은 점수로 평가 받았다. 이를 통해 연구에서는 폐쇄망 조건을 유지하면서 근거성 있는 질의응답을 가능하게 하는 실용 아키텍처를 제시하였으며, 이후 한-영 동시 온톨로지 확장과 이중언어 색인 및 하이브리드 검색, 프롬프트 최적화 등을 통해 폐쇄망에서의 LLM 시스템 구현에 보탬이 되고자 한다.

1. 서론

대규모 언어모델(이하 LLM)은 다양한 도메인 기반의 문서를 근거로 하는 질의 응답을 가능하게 하지만, 일반적으로는 클라우드 자원이나 외부 네트워크에 의존하기 마련이다. 전력이나 국방, 의료 등 폐쇄망이 요구되는 환경에서는 기밀성과 가용성이 요구되므로 외부 접속이 불가능하거나 엄격히 제한되기 때문에 오프라인에서 동작하는 문서 검색 및 질의응답 체계가 필요하다.

따라서 본 연구를 통해 제안하는 시스템은 첫 번째, 도메인 약어/라벨을 자동으로 채굴하여 온톨로지를 생성하는 전처리 단계, 두 번째 온톨로지로 질의를 확장하고 개념 적합도를 반영하는 재랭킹 기반 로컬 검색 단계, 세 번째, 완전 오프라인 LLM 추론 단계, 네 번째, 로컬 전용 UI를 통한 운용 편의성으로 구성된다.

전처리 단계에서는 PDF/TXT에서 헤더(Header) 및 푸터(footer) 문단 및 문장 단위의 청킹, 키워드 추출을 수행하고 약어-라벨 패턴을 감지하여 'Ontology.Json'을 자동 생성함으로써 이후 검색 개념 맥락을 강화한다.

검색 및 추론 단계에서는 온톨로지 기반 질의 확장과 개념 적합도 재랭킹을 결합하여, 도메인 약어와 동의어가 혼재한 문서에서도 관련 청크를 안정적으로 선별한다. 이후 로드한 GGUF 모델이 오프라인에서 응답을 생성하므로 폐쇄망 정책을 충족하면서도 RAG 절차의 이점을 확보할 수 있다.

연구는 우선, 도메인 문서에서 약어-라벨을 자동 채굴해 온톨로지를 생성 및 갱신하는 전처리 파이프라인을 제시하며, 온톨로지 확장 질의와 개념 적합도 재랭킹을 결합한 로컬 RAG 인덱싱 검색 기법을 설계한다. 또한, GGUF 모델을 활용하여 오프라인 추론을 구현하고, 폐쇄망에 맞는 UI 및 배포 설정을 정립한다.

2. 관련 연구

2.1 LLM 오픈소스 모델의 확산

LLM의 개방과 경량화는 온프레미스 및 폐쇄망 환경에서의 실사용 가능성을 높였다. Meta의 Llama 2는 7B–70B 파라미터 범위의 사전학습/대화형 모델을 공개하여, 상용 폐쇄형 모델에 근접한 품질을 오픈 모델로 구현할 수 있음을 확인하였다[1].

2.2 RAG: 생성과 검색의 결합

도메인 지식의 최신성과 근거성, 갱신 용이성을 확보하기 위해 생성 모델과 외부 비매개 메모리를 결합한 RAG(Retrieval–Augmented Generation)가 표준 설계로 자리 잡았다. rag는 시퀀스–투–시퀀스 생성기에 외부 색인을 조합해 사실성과 특이성 등을 개선한다.

성능은 검색 품질에 좌우되며, 전통적 희소 모형인 BM25 등은 단순 해석 가능성이 강점으로, 확률적 관련성 프레임워크로 체계화되어 왔다. 반면, Dense Passage Retrieval(DPR)은 질의와 문서의 밀집 임베딩으로 대규모 오픈 도메인 QA에서 BM25 대비 상위 문서 적중률을 유의하게 끌어올렸다[2].

2.3 온톨로지·지식그래프 기반 검색 고도화

온톨로지와 지식그래프를 활용한 쿼리 확장, 그리고 Semantic 검색은 용어 및 개념 불일치 문제를 줄이고, 다중 hop 추론이 필요한 질의에서 맥락을 보강한다는 점이 선행 연구에서 정리되어 왔다.

최근에는 KG/온톨로지를 RAG 파이프라인에 직접 결합해 색인 구조 및 경로 탐색, 근거 귀속을 개선하려는 시도가 활발하며, 생의학 QA 등 전문 영역에서 KG–보강 RAG가 복잡 질의에 효과적이라는 실증도 보고되었다. 연구에서는 이러한 흐름을 폐쇄망 도메인으로 확장하여, 온톨로지 기반 개념 정규화와 관계 추론을 로컬 LLM–RAG에 결합하는 설계를 다루고자 하였다[3].

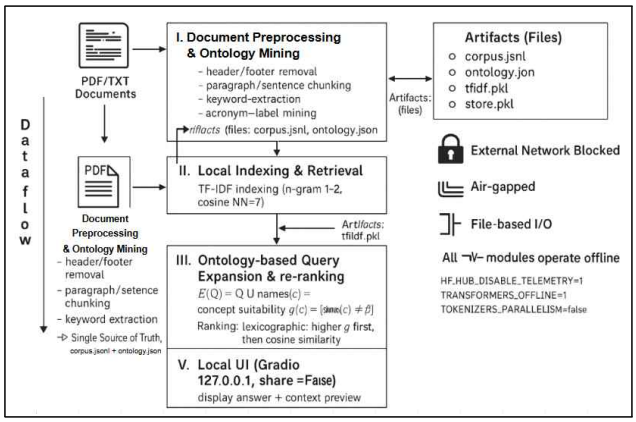
2.4 국내 연구 동향

국내에서는 온톨로지 기반 정보검색 시스템 설계와 구현에 관한 연구가 다수 보고되어 왔고, 최근에는 RAG의 최신 동향 및 적용 과제에 대한 종설도 발표되었다. 이러한 점은 도메인 개념 모델을 활용한 질의 정규화·근거 제시·환각 저감의 필요성을 강조한다고 볼 수 있다[4,5].

3. 시스템 아키텍처와 구성요소

3.1 전체 아키텍처

시스템은 폐쇄망 환경에서 동작하는 온톨로지–강화 로컬 RAG 및 LLM 파이프라인으로, [그림1]과 같이 작동한다.



[그림 1] 시스템 아키텍처 개요

3.2 전처리·온톨로지 모듈

전처리는 PDF/TXT에서 텍스트 추출 후 반복되는 헤더 및 푸터를 통계적으로 제거하고, 문장·문단 단위 청킹을 통해 검색 친화적 단위로 정규화 한다. 동시에 약어–라벨 패턴을 정규식으로 채굴하여 누적 온톨로지 버퍼에 병합하고, 도메인 시드와 통합해 ontology.json 파일을 생성한다. 수식적으로 문서 d의 문장열을 청킹 함수 $C(\cdot)$로 분할하면 다음과 같고,

$$C(d) = c_1, \ldots, c_m, \ | c_i | \in [\text{min\_words}, \ \text{max\_words}],$$

약어–라벨 정규식 R로부터 발견된 개념 집합을 O라 할 때 최종 온톨로지는 아래 수식과 같이 정의된다.

$$ONTO = freeze\left(SEEDS \cup \bigcup_{i=1}^{m} R(c_i)\right)$$

3.3 로컬 인덱싱·검색 모듈

인덱싱은 TF–IDF 기반 희소 표현을 사용하며, 각 청크 c에 대해 용어 t의 가중치는 다음 수식과 같다.

$$tfidf(t,c) = tf(t,c) \cdot \log \frac{N+1}{df(t)+1}$$

문서 벡터는 L_2 정규화된 V_C로 표현한다. 질의 q는 3.4절의 온톨로지 확장 후 벡터화하여 코사인 거리를 통한 최근접 탐색으로 후보 상위 K개를 수집한다.

3.4 온톨로지 기반 질의 확장

온톨로지에서 개념 c의 명칭 집합(약어·라벨·별칭)을 $names(c)$라 하면, 원 질의 토큰 집합 Q에 대해 확장 질의 $E(Q)$는 다음 수식과 같다.

$$E(Q) = Q \cup \bigcup_{c \in C(Q)} names(c),$$

$$C(Q) = c \in ONTO vert names(c) \cap Q \neq \varnothing$$

즉, 질의에 등장한 도메인 약어/라벨을 기준으로 동의어·별칭을 자동 결합하여 용어-개념 불일치로 인한 누락을 줄인다.

3.5 개념 적합도 재랭킹

검색 단계에서는 코사인 유사도를 기반으로 상위 문서를 선정함에 그치지 않고, 온톨로지 기반의 개념 적합도를 추가로 고려한다.

질의어와 관련된 개념이 다수 포함된 청크는 도메인 적합성이 높다고 판단하며, 이런 청크를 상위에 배치한다. 이후 개념 적합도가 동일한 경우에는 코사인 유사도를 보조 기준으로 삼아 순위를 결정한다.

3.6 로컬 LLM 추론 모듈

검색 및 재랭킹을 통해 선정된 상위 청크는 로컬 LLM의 추론 모듈로 전달된다. 연구에서는 llama.cpp를 활용한 GGUF 형식 모델을 사용하였으며, 모든 추론 과정은 오프라인에서 이뤄진다.

프롬프트 구조는 시스템 규칙, 검색된 컨텍스트 매칭된 개념 목록, 사용자의 질의로 구성되며, 시스템 규칙에는 답변이 반드시 컨텍스트에 기반해야 하며, 최대 6문장 이내 제한, 마지막 문장은 반드시 완결된 문장으로 종료해야 한다는 지침을 포함하였다.

4. 실험 결과

4.1 실험 구성

실험은 폐쇄망으로 구성된 로컬 워크스테이션에서 수행하였으며, 전처리는 PDF/TXT 문서를 입력받아 헤더/푸터 제거-문단/문장 청킹-키워드 추출-약어/라벨 채굴을 거쳐 corpus.jsonl과 ontology.json을 생성한다. 출력물은 후속 단계의 단일 SSOT로 사용된다.

인덱싱 검색기는 TfidVectorizer와 NearestNeighbors(cosine)로 희소 인덱스를 구축하고, 질의 시 온톨로지 기반 확장과 개념 적합도 우선 정렬을 적용한다. 이후 선택된 컨텍스트는 llama.cpp 기반 GGUF 모델로 전달되어 오프라인에서 답변을 생성한다. 질의-검색-생성 파이프라인은 Gradio UI로 제공하였다.

4.2 데이터셋 및 과제

평가 질의는 각 질의에 대해 네트워크에 연결된 Chat GPT와 로컬 GPT를 실행하여 각각 비교하였다. 두 시스템은 동일한 사용자 지침과 동일 질의를 사용하였고, Chat GPT의 경우 외부 검색 기능은 사용하지 않았으며, 로컬 GPT는 제안 파이프라인의 RAG 컨텍스트를 사용하였다. 결과는 정답 일치(EM/F1), 근거 포함 여부(Evidence Hit Rate), 지연시간(End-toEnd Latency)으로 집계하였다.

실험에 이용된 임베딩 코퍼스와 대표 질의 예시는 다음 [표 1]과 같다.

[표 1] 임베딩 코퍼스 및 질의 예시

Source Doc(언어)	핵심 범위/용어	질의 예시
KOSHA Guide E-185-2021 (KO)	리튬이온 ESS 설치·운영·방호, BMS/PCS/EMS 정의 등	컨테이너형 ESS의 설치 위치와 화재 방호 기준은?
NISTIR 7628(EN)	스마트그리드 보안 전략·아키텍처·인터페이스	NISTIR 7628의 7개 도메인과 일반적인 인터페이스 범주를 나열해주세요.
송·배전용 전기설비 이용규정(KO)	접속·보호협조·계통연계 절차/요건, 서식	배전 접속 시 보호협조 핵심 요건은?
ISGAN White Paper: Smart Grid Cyber Security (EN)	정책·조직 이슈, Defense-in-Depth, 프라이버시 프레임워크, 5대 보안 성질	5가지 핵심 보안 속성을 무엇이며, 애플리케이션에 따라 왜 다른가요?

4.3 평가 결과

동일 질의셋에 대해 LocalGPT와 온라인 Chat GPT의 응답을 비교하였다. 두 시스템의 각 질의 응답 쌍에 대해 먼저 Similarity Score로 두 응답의 의미적 유사도를 인간 평가자가 0-5점으로 채점하고, Processing Time으로 로컬 파이프라인의 end-to-end 지연을 ms 단위로 기록하였다. 비고에는 각 질의에서 사용된 근거 인덱싱 문서의 주언어(KO/EN)를 표시하였다. 평가자는 2인으로 구성되었으며, 평균 점수를 차용하였고, 그 결과는 아래 [표 2]과 같다.

[표 2] 'LocalGPT', 'Chat GPT 4.o' 비교 (일부)

Query ID	Similarity Score	Processing Time (ms, LocalGPT)	Index Doc Lang
Q-01	3.5	2180	KO
Q-02	2.0	2335	EN
Q-03	1.0	2510	EN
Q-04	5.0	1970	KO
Q-05	4.0	2250	KO
Q-06	3.0	2395	EN
Q-07	4.0	2110	KO
Q-08	0.0	2290	EN
Mean±Std	2.81±1.69	2255±169	-

전체 평균 유사도는 2.81(±1.69)점으로 편차가 존재하였으며, 질의별로 Q-03, Q-08과 같이 두 시스템의 답변이 갈리는 사례가 존재하였다. 지연시간은 평균 2,255ms로 비교적 안정적이었으나 질의 난이도와 직접적인 상관은 크지 않았다.

언어별로는 한국어 기반의 질의에서 유사도 4.13점(±0.63)으로 안정적이었으나, 영어 기반 질의에서는 유사도가 떨어지는 현상이 발생하여 평균 약 1.5점에 머물렀고, 지연시간도 약 2,383ms로 한국어 기반 질의에 비해 길었다. 이는 인덱싱 코퍼스가 한국어 중심이고, 온톨로지의 영어 동의어/별칭 커버리지가 제한적이라 질의 확

장부터 검색, 컨텍스트 선정까지의 과정이 약화된 결과로 해석된다.

종합하자면, 먼저 온톨로지의 영어, 한국어 동시 시드 확장(라벨·별칭·약어), 다음으로 영어 질의의 사전 번역·용어 정규화, 마지막으로 이중언어 인덱스 및 문자·서브워드 n-gram 도입을 통해 검색과 컨텍스트 품질을 보강하는 것이 효과적일 것으로 보여진다.

5. 결론

연구에서는 폐쇄망 환경을 전제로 온톨로지 기반 질의 확장과 개념 적합도 재랭킹을 결합한 로컬 RAG+LLM 파이프라인을 설계 및 구현하였다. 시스템은 첫 번째, 약어·라벨 채굴을 포함한 전처리로 corpus.jsonl과 ontology.json을 생성하고, 두 번째로 TF-IDF 인덱싱 위에 온톨로지 기반 질의 확장과 개념 적합도 우선 정렬을 적용한 컨텍스트를 선별하며, 세 번째로 llama.cpp 기반 GGUF 모델로 오프라인 추론을 수행하였다.

실험에서는 동일 질의셋에서 LocalGPT와 ChatGPT에 Context를 결합한 모델을 비교하여 인간 평가를 진행하였고, 평균은 2.81점으로(±1.69) 나타났다. 로컬 파이프라인의 처리 지연은 2,255ms(±169ms)로 안정적이었고, 언어별로 한국어 질의에서 유사도가 4.13점(±0.63)으로 높고 지연도 더 짧았으나, 영어 기반 질의에서는 유사도가 낮아지는 사례가 다수 관찰되었다. 이는 인덱싱 코퍼스의 한국어 편중과 온톨로지의 영어 동의어·별칭 커버리지 한계가 결합해, 질의 확장부터 검색, 컨텍스트 품질까지의 과정이 약화된 것에 기인한 것으로 해석된다. 그럼에도 불구하고, 시스템은 보안과 운영 제약을 만족하면서 동시에 한국어 기반 업무 질의에 대해 클라우드 모델과 근접한 의미적 일치를 달성함을 확인하였다.

본 연구의 주요 기여는 다음과 같다.

첫 번째로, 온톨로지 자동화 파이프라인을 통해 약어-라벨을 추출·정규화하고, 전처리 산출물을 단일 진실원(SSOT)으로 표준화 하였다.

두 번째로, 개념 적합도 재랭킹을 도입하여, 전문 용어가 밀집한 문서에서 컨텍스트 선별의 안정성을 높였다.

세 번째, 오프라인 LLM 추론 및 로컬 UI를 통해 폐쇄망 환경의 실사용 가능성을 입증하였다. 한편 한계도 존재하였다. 희소기반 검색은 용어 변형과 의미 유사성에 취약하며, 온톨로지의 영어 커버리지가 제한되어 영어 기반 질문에서는 성능 편차가 높았으며, 평가 규모가 크지 않고 사람 평가 중심이라 통계적 일반화에 제약이 존재한다.

이에 따라 향후 영문 라벨과 별칭, 약어를 시드로 확장하고 정규식 패턴 빈도 통계 결합으로 자동 병합 주기를 단축하는 방향과, 한국어, 영어 동시 색인, 문자·서브워드 n-gram, BM25+Dense 혼합으로 영문 질의의 컨텍스트 품질을 보강하는 연구를 제안한다.

참고문헌

[1] P. Lewis 외, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”, Advances in Neural Information Processing Systems, Vol. 33, pp. 9459-9474, 12월, 2020년. DOI: 10.48550/arXiv.2005.11401.

[2] S. Robertson, H. Zaragoza, “The Probabilistic Relevance Framework: BM25 and Beyond”, Foundations and Trends in Information Retrieval, Vol. 3, No. 4, pp. 333-389, 2009년. DOI: 10.1561/1500000019.

[3] N. Matsumoto 외, “KRAGEN: a knowledge graph-enhanced RAG framework for biomedical problem solving using large language models”, Bioinformatics, 40(6): btae353, 2024년. DOI: 10.1093/bioinformatics/btae353.

[4] 윤여찬, 김수균, “생성형 AI를 위한 Retrieval-Augmented Generation (RAG) 기술 동향 및 전망,” 한국컴퓨터교육학회 논문지, 제28권 제2호, pp. 7-26, 2월, 2025년. DOI: 10.32431/kace.2025.28.2.007.

[5] 김도현, 원일웅, 유상현, 김현정, “RAG 시스템 성능 향상을 위한 웹문서 본문 정제 및 추출 시스템,” 한국통신학회 인공지능 학술대회 논문집, pp. 305-306, 9월, 2024년.