

초거대 언어 모델과 다중 에이전트를 활용한 실시간 대기질 데이터 분석 시스템

지건우, 지영민, 권동우

한국전자기술연구원

geonwoo24@keti.re.kr, ym.ji@keti.re.kr, dwkwon@keti.re.kr

Real-time Air Quality Analysis System with Large Language Model and Multi-Agent

Geonwoo Ji, Youngmin Ji, Dongwoo Kwon

Korea Electronics Technology Institute

요약

대기오염은 인류 건강에 심각한 위협을 가하는 주요 환경 문제로, 실시간 대기질 데이터를 정확히 분석하고 효과적으로 활용하는 것이 필수적이다. 그러나 방대한 양의 대기질 데이터를 처리하고 분석하는 데에는 상당한 시간과 전문 인력이 요구되며, 이는 실시간 대응과 효율적인 의사결정을 어렵게 만든다. 이에 본 연구는 초거대 언어 모델 (large language model, LLM) 기반의 다중 에이전트 시스템을 활용하여 대기질 데이터를 효율적으로 분석하고, 보고서를 자동으로 생성하는 시스템을 제안한다. 제안하는 시스템은 대기질 데이터의 탐색, 분석, 상황 인식, 보고서 작성 등 전 과정을 역할 기반 에이전트 간의 협업을 통해 수행함으로써, 기존 수작업 방식 대비 분석에 소요되는 시간과 인적 자원 부담을 효과적으로 줄일 수 있다. 또한, 에어코리아의 실시간 대기질 데이터를 활용한 분석 시나리오를 통해 시스템의 유효성을 검증하였으며, 제안하는 시스템이 대기환경 분석 및 모니터링과 같이 실시간성과 자동화가 요구되는 업무에서 전문가의 의사결정을 효과적으로 보조할 수 있을 것으로 기대한다.

I. 서론

대기오염은 세계보건기구(WHO)가 매년 약 700만 명의 조기 사망을 유발하는 주요 원인으로 지목할 만큼 심각한 사회적 문제이다. 초미세먼지(PM_{2.5}), 오존(O₃), 이산화질소(NO₂), 이산화황(SO₂), 일산화탄소(CO) 등 주요 대기오염물질은 심혈관 및 호흡기 질환은 물론, 뇌혈관계 질환과 암 등 다양한 질병의 발병 위험을 증가시키며, 특히 어린이, 노인, 만성질환자와 같은 취약계층에게 치명적인 영향을 미친다. 따라서 대기오염 상태를 정확히 파악하고 인체에 미치는 악영향을 최소화하기 위해서는 대기질 정보를 지속적으로 분석하고 모니터링하는 것이 필수적이다. 하지만 실시간으로 수집되는 방대한 양의 대기질 데이터는 사람이 직접 분석하고 이를 기반으로 의사결정을 내리기까지 상당한 시간과 전문 인력이 요구된다.

최근 초거대 언어 모델 (large language model, LLM)의 성능이 비약적으로 향상되면서, 검색 증강 생성 (retrieval-augmented generation, RAG)과 다중 에이전트 (multi-agent) 기반 기술이 결합되어 데이터 해석, 요약, 보고서 작성 및 작업 자동화 등 다양한 영역에서 혁신을 이끌고 있다 [1][2]. 이러한 기술들은 여러 산업 전반에 걸쳐 적용 가능성이 주목되고 있으며, 관련 연구도 활발히 진행되고 있다.

Kim과 Shin은 전력 산업에서 LLM을 활용하기 위한 포괄적인 분석과 전략을 제시하였고, 특히 보안성과 제도적 제약을 고려한 에너지 분야 특화 LLM의 활용 방안을 구체적으로 제안하였다 [3].

Xinxin 등은 LLM과 계층적 코퍼스를 결합한 SRAG 프레임워크를 제안하여, 요약 보고서를 자동으로 생성하고 다단계 절차를 통해 보고서의 구조화, 정확성, 완성도를 향상시켰다 [4].

또한 Chang 등은 의료 영상 보고서의 자동 생성을 위해 LLM을 의료 도메인에 특화시키는 학습 기법을 개발하였으며, 도메인 적응과 정밀 생성의 두 단계를 통해 기존 방법보다 뛰어난 성능을 입증하였다 [5].

이러한 기술 발전은 대기환경 분야에서도 새로운 가능성을 제시한다. 기존에는 도메인 전문가가 수작업으로 수행하던 데이터 탐색, 분석, 상황 판단, 보고서 작성 등의 과정을 LLM 기반 다중 에이전트가 실시간으로 처리하여 분석의 효율성과 정확도를 동시에 향상시킬 수 있다. 이는 단순한

작업 자동화를 넘어, 전문가의 의사결정을 지원하고 강화하는 방향으로 발전 가능하다. 하지만 사실과 다른 내용을 사실처럼 생성하는 환각 (hallucination) 문제나 실시간 데이터의 접근 어려움 등 해결해야 할 기술적 과제가 여전히 존재한다. 이에 본 논문에서는 초거대 언어 모델 기반 다중 에이전트 시스템을 활용하여, 대기질 데이터 분석에 필요한 시간적, 인적 자원 부담을 효과적으로 줄일 수 있는 실시간 대기질 데이터 분석 시스템을 제안한다. 제안하는 시스템은 검색 증강 생성, 도구 호출, 다중 에이전트 협업을 통해 환각과 같은 오류 발생 가능성을 줄이고, 실시간 데이터 접근을 가능하게 한다. 또한, 제안한 시스템의 유효성을 검증하기 위해 시나리오를 설계하고, 전국 663개 대기질 측정소로부터 실시간으로 수집되는 공공 데이터인 에어코리아 대기질 데이터를 활용하여 실험을 수행하였다.

II. LLM 및 다중 에이전트 기반 실시간 대기질 데이터 분석 시스템

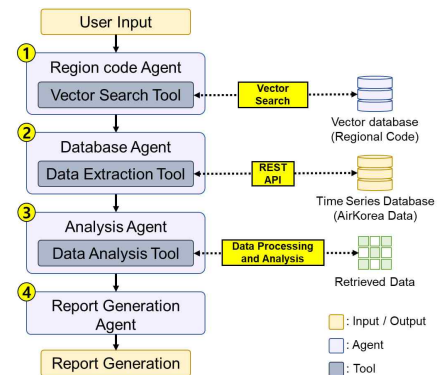


그림 1 다중 에이전트 기반 실시간 대기질 데이터 분석 시스템 구조

그림 1은 제안하는 LLM 및 다중 에이전트 기반 실시간 대기질 데이터 분석 시스템의 전체 구조를 나타낸다. 본 시스템은 여러 에이전트가 상호

협업을 통해 복합적인 작업을 수행하는 다중 에이전트로 구성되어 있으며, 총 네 개의 주요 에이전트가 고유한 기능을 담당하여 전체 분석 프로세스를 구성한다. 이러한 구조를 통해 단일 에이전트를 활용하는 방식보다, 역할 분담과 적절한 도구의 활용을 통해 분석의 정확도를 효과적으로 향상시킬 수 있다. 모든 에이전트는 META사의 LLaMA 3 모델을 기반으로 하여 구현되었다 [6].

우선, 사용자가 “성남시의 최근 일주일 대기질을 분석해줘”와 같은 질의를 입력하면, 해당 요청은 첫 번째 단계에서 지역 코드 에이전트에게 전달된다. 에어코리아 대기질 데이터에 접근하기 위해서는 분석 대상 지역의 고유 코드가 필수적이며, 이를 위해 지역 코드 에이전트는 추출 도구를 호출하여 질의를 벡터 형태로 변환한 뒤, 유사도 검색을 통해 벡터 데이터베이스(Vector DB)로부터 가장 관련성 높은 지역 코드를 추출한다. 이 과정은 단순 LLM 기반 접근 방식에서 발생할 수 있는 환각 문제를 완화하고, 보다 정확한 정보 분석을 가능하게 한다.

다음 단계에서는 지역 코드와 원 질의가 함께 데이터베이스 에이전트에게 전달된다. 본 시스템은 검증된 대기질 데이터인 에어코리아 데이터를 별도로 구축한 시계열 데이터베이스에 저장하여 활용하였으며, 이 에이전트는 데이터 조회 도구를 사용하여, 에어코리아 대기질 데이터가 저장되어있는 시계열 데이터베이스로부터 관련 데이터를 조회한다. 데이터 조회는 REST API를 통해 이루어지며, 이 과정에서 에이전트는 입력된 지역 코드 및 원 질의를 기반으로 데이터베이스의 매트릭-태그 스키마에 대한 매핑을 수행하고, 이를 바탕으로 API 요청을 구성한다.

조회된 데이터는 분석 에이전트에게 전달되며, 해당 에이전트는 분석 도구를 호출하여 데이터를 정제하고 분석한다. 분석 도구는 최대값, 최소값, 일평균 변화율 등 주요 통계 기반 지표를 산출하며, 모듈화되어 있기 때문에 분석 목적에 따라 다양한 방식으로 유연하게 확장 및 통합이 가능하다.

마지막으로, 분석 결과는 보고서 작성 에이전트에게 전달된다. 보고서 작성 에이전트는 분석된 정보를 종합하고, 도메인 지식 및 문맥적 정보를 반영하여 최종 분석 보고서를 자동 생성한다. 생성된 보고서는 사용자의 초기 질의와 관련 데이터에 기반한 정량적 분석 결과와 함께, 해석 가능한 정성적 설명을 포함하여 구성된다.

III. 에어코리아 대기질 데이터 분석 시나리오

제한한 시스템의 유효성을 검증하기 위해 대기질 데이터 분석 시나리오를 설계하고 실험을 수행하였다. 실험 시나리오는 2025년 5월 1일부터 5월 7일까지 최근 일주일간의 성남시 대기질 데이터를 분석하는 상황을 가정하였다. 해당 시나리오에 따라 분석 시스템에 요청을 입력한 후, 시스템의 응답 결과를 분석하였고, 응답 결과는 그림 2에 보이는 바와 같다.

출력된 보고서는 실시간 에어코리아 대기질 데이터를 활용하여 명확하게 작성되었으며, 각 유해 인자별 수치 변화가 정량적으로 잘 정리되었다. 또한 정량적 분석 결과와 도메인 지식, 문맥 정보를 반영하여 작성된 종합 분석 내용은 전체적인 대기 환경 상태를 잘 설명하고 있으며, 필요한 후속 조치에 대해서도 구체적으로 제시하고 있다. 한편, 작성된 보고서의 ‘분석 결과’, ‘외부 활동 여부’ 항목에서는 간혹 한자가 혼용된 표현이 나타났다. 이는 시스템이 기반으로 사용하는 LLaMA 3 언어 모델의 특성과 관련이 있다. 해당 모델은 영어와 한국어를 포함한 다국어 데이터를 학습하였으며, 이 과정에서 일부 한국어 문서에 특정 단어가 한자로 표기된 사례가 포함되었을 가능성이 있다. 이러한 현상은 토큰나이저 처리 방식 등 다국어 모델에서 흔히 발생하는 문제로, 한국어 전용 언어 모델의 활용, 프롬프트 조정, 출력 후처리, 또는 언어 모델의 미세 조정(fine-tuning)을 통해 개선할 수 있다.

2025년 5월 1일부터 7일간의 성남시 대기 환경 분석 보고서

1. 제목 및 기본 정보

- 지역명: 성남시
- 분석 기간: 2025년 5월 1일 ~ 2025년 5월 7일

2. 요약 (Executive Summary) 성남시의 대기 환경 분석 결과, 미세먼지와 초미세먼지의 농도는 상대적으로 높은 편으로 확인되었습니다. 일산화탄소, 오존, 이산화황, 이산화질소의 농도는 비교적 낮은 편이지만, 지속적인 모니터링이 필요합니다. 대기 환경 상태는 중간 수준으로 평가되며, 특히 미세먼지와 초미세먼지에 대한 주의가 필요함을 시사합니다.

3. 대기환경청서값 분석 내용

- 미세먼지(PM10): 최대값 114.0 μ g/m³, 최소값 12.0 μ g/m³, 평균값 38.905 μ g/m³, 일평균 변화율 1.4%
- 초미세먼지(PM2.5): 최대값 78.0 μ g/m³, 최소값 2.0 μ g/m³, 평균값 15.633 μ g/m³, 일평균 변화율 4.3%
- 일산화탄소(CO): 최대값 0.6ppm, 최소값 0.2ppm, 평균값 0.353ppm, 일평균 변화율 0.9%
- 오존(O3): 최대값 0.077ppb, 최소값 0.01ppb, 평균값 0.043ppb, 일평균 변화율 1.8%
- 이산화황(SO2): 최대값 0.003ppb, 최소값 0.001ppb, 평균값 0.002ppb, 일평균 변화율 1.7%
- 이산화질소(NO2): 최대값 0.034ppb, 최소값 0.005ppb, 평균값 0.014ppb, 일평균 변화율 1.4%

미세먼지와 초미세먼지의 농도는 상대적으로 높은 편으로, 대기 오염에 대한 주의가 필요합니다. 일산화탄소, 오존, 이산화황, 이산화질소의 농도는 비교적 낮은 편이지만, 지속적인 모니터링이 필요합니다.

4. 종합 분석 내용 정리

- 분석 결과: 성남시의 대기 환경 상태는 중간 수준으로 평가되며, 특히 미세먼지와 초미세먼지에 대한 주의가 필요함을 시사합니다.
- 외부 활동 여부: 외부활동時 미세먼지 및 초미세먼지에 대한 주의가 필요하며, 마스크 착용 등 예방책을 권장합니다.
- 대처 방안: 대기 오염에 대한 주의와 예방책을 강조하며, 추가적인 연구와 모니터링이 필요합니다. 시민들에게 대기 환경 정보를 제공하고, 대기 오염 예방을 위한 교육 및 캠페인을 실시할 수 있습니다.

이 분석 결과는 성남시의 대기 환경 정책과 관리에 도움이 될 것입니다. 또한 시민들에게 대기 오염에 대한 주의와 예방책을 강조할 수 있을 것입니다.

그림 2 대기질 데이터 분석 시나리오 실험 결과 생성된 보고서

IV. 결론

본 논문에서는 초거대 언어모델 기반의 다중 에이전트 시스템을 활용한 실시간 대기질 데이터 분석 시스템을 제안하고, 에어코리아 대기질 데이터를 활용한 분석 시나리오를 통해 그 유효성을 검증하였다. 제안된 시스템은 데이터 탐색, 분석, 상황 판단, 보고서 작성 등 주요 분석 과정을 각기 다른 역할의 에이전트들이 협업하여 수행함으로써, 기존의 수작업 중심 분석 방식에 비해 시간적, 인적 자원 부담을 크게 줄일 수 있음을 확인하였다. 또한 본 시스템은, 실시간성과 자동화 수준이 요구되는 대기환경 분석 및 모니터링 업무에서 전문가의 의사결정을 효과적으로 보조하고, 정량적, 정성적 분석의 정확도 향상에 기여할 수 있을 것으로 기대된다. 향후에는 다양한 기상 변수와 다종 데이터를 연계한 통합 분석으로 확장함으로써, 대기오염 대응 전략 수립에 더욱 효과적인 도구로 발전시킬 수 있을 것이다.

ACKNOWLEDGMENT

본 연구는 산업통상자원부(MOTIE)와 한국에너지기술연구원(KETEP)의 지원을 받아 수행한 연구 과제입니다. (RS-2020-KP002220)

참 고 문 헌

- [1] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Advances in Neural Information Processing Systems, vol. 33, pp. 9459 - 9474, 2020.
- [2] Y. Talebirad and A. Nadiri, "Multi-agent collaboration: Harnessing the power of intelligent LLM agents," arXiv preprint, arXiv:2306.03314, 2023.
- [3] E. Kim and Y. Shin, "A study on the direction of large language model (LLM) utilization in the domestic power industry," J. Korean Inst. Commun. Inf. Sci., vol. 50, no. 2, pp. 217 - 223, 2025.
- [4] G. Xinxin, et al. "Summary report auto-generation based on hierarchical corpus using large language model," Displays (2025): 103055.
- [5] L. Chang, et al. "Bootstrapping large language models for radiology report generation," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 17, 2024.
- [6] A. Dubey et al., "The Llama 3 Herd of Models," arXiv preprint, arXiv:2407.21783, 2024.