LLM 기반 멀티 에이전트 시스템에서 토폴로지 및 언어에 따른 정보 유출 분석*

백가현[O], 김민석, 구형준
성균관대학교 소프트웨어학과
gahyun2097@g.skku.edu, for8821@g.skku.edu, kevin.koo@skku.edu

A Study of Information Leakage Across Topologies and Languages in LLM-based Multi-Agent Systems

Gahyun Baek[O], Minseok Kim, Hyungjoon Koo
Department of Computer Science and Engineering, Sungkyunkwan University

요 약

최근 대규모 언어 모델 (LLM)은 AI 에이전트가 협력할 수 있는 멀티 에이전트 시스템 (MAS)을 지원해 높은 효율성과 확장성을 제공한다. 한편, 에이전트 간 통신 메시지나 공유 메모리를 통한 정보 유출이 가능해지면서 새로운 보안 위협을 수반한다. 본 연구는 기존의 에이전트 프레임워크에서 두 가지 언어와 세 가지 네트워크 토폴로지 (단일 계층, 탈중앙화, 다중 계층)에서 정보 유출 양상을 분석했다. 실험 결과, 언어 환경에 관계없이 탈중앙화 구조가 다른 구조 (단일 계층, 단일 계층)에 비해 평균적으로 약 21.5% 더 높은 유출률을 보이며 가장 취약한 것으로 나타났다. 또한, 두 언어 환경을 비교했을 때 국문 환경이 영문 환경 대비 8.7% 더 많은 정보 유출이 발생함을 확인하였다.

1. 서 론

최근 대규모 언어 모델 (Large Language Model, LLM)은 단순한 대화형 인터페이스를 넘어, 외부 환경과 능동적으로 상호작용하며 자율적으로 목표를 달성하는 AI 에이전트 (Agent) 형태로 진화하고 있다 [1]. 나아가 복수의 에이전트가 복잡한 작업을 역할별로 분산·위임하여 협력하는 멀티 에이전트 시스템 (Multi-Agent System, MAS)은 기존 단일 에이전트 시스템 대비 높은 효율성과 확장성을 보이며, 다양한 산업 분야에서 자동화된 의사결정 및 정보 처리 시스템으로 활용되고 있다. 그러나 이러한 구조적 이점은 보안 측면에서 기존 환경과는 상이한 새로운 공격 표면을 수반한다.

멀티 에이전트 시스템의 보안을 다루는 선행 연구로 AgentDojo [2]와 AirGapAgent [3] 등이 있으나, 이들은 주로 단일 에이전트 수준의 위협에 초점을 맞추고 있어 다중 에이전트 간 상호작용에서 발생하는 내부 정보 유출에 대한 분석은 미흡하다. 반면, AgentLeak [4]은 최종 출력뿐 아니라 에이전트 간 메시지, 공유 메모리, 외부 도구 상호작용을 포함한 시스템 전반의 정보 유출을 평가하는 벤치마크를 제시했다. 다만 해당 연구는 단일 토폴로지 (Topology)와 단일 언어 환경 (영어)에 한정되어 실험이 수행되었다는 한계가 있다.

본 논문에서는 이러한 한계를 보완하고자 AgentLeak을 기반으로, 실험 환경을 한국어 생태계 및 다양한 토폴로지 구조로 확장하여 정보 유출 양상을 평가한다. 분석 결과, 언어 환경과 무관하게 탈중앙화 구조가 다른 구조 (다중 계층, 단일 계층)에 비해 평균적으로 약 21.5% 더 높은 유출률을 보이며 가장 취약한 것으로 나타났다. 나아가 언어별 차이를 비교했을 때, 국문 환경에서 에이전트 간 메시지를 통한 정보 노출량이 영어 환경에서보다 약 8.7% 더 높게 나타남을 확인하였다.

* 본 논문은 2026년도 정부 (과학기술정보통신부) 한국연구재단 (생성모델 컴플라이언스를 위한 모듈형 AI 워터마킹 기술 연구실; No.RS-2025-02293072) 지원으로 수행한 연구임.

2. 배경지식

AI 에이전트. AI 에이전트는 대규모 언어 모델을 핵심 추론 엔진으로 활용하여, 주변 환경을 인지하고 스스로 의사결정을 내리며 행동을 수행할 수 있는 인공 객체이다 [5]. 기존의 대화형 LLM (예: ChatGPT, Claude 등)이 사용자 질의에 대해 수동적으로 응답을 생성하는 데 그쳤다면, AI 에이전트는 목표 지향적 계획 수립 (planning), 외부 환경과의 상호작용 (tool use), 그리고 실행 결과에 기반한 반복적 추론 (iterative reasoning) 과정을 통해 복합적인 문제를 능동적으로 해결한다.

AgentLeak. AgentLeak [4] 은 멀티 에이전트 시스템에서 발생하는 정보 유출 (예: 개인정보)을 시스템 전반에 걸쳐 포괄적으로 평가하는 벤치마크이다. 기존 연구가 최종 출력 중심으로 정보 노출을 분석한 것과 달리, 에이전트 간 내부 통신과 공유 메모리 및 외부 도구 상호작용까지 평가 범위에 포함함으로써 멀티 에이전트 환경의 유출 위험을 보다 포괄적으로 다룬다. 다만, 해당 연구는 영어로만 진행했다는 점과 단일 계층 구조의 토폴로지에만 한정되었다는 한계를 지닌다. 이에 본 연구에서는 AgentLeak의 실험 범위를 확장하여, 두 가지 토폴로지 (탈중앙화 구조, 다중 계층형 구조)와 한국어 환경을 추가함으로써 토폴로지 및 언어가 정보 유출에 미치는 영향을 분석한다.

3. 토폴로지 및 언어에 따른 정보 유출 분석

3.1 위협 모델 및 공격 표면

그림 1에 나타낸 바와 같이, 본 연구는 AgentLeak [4]의 위협

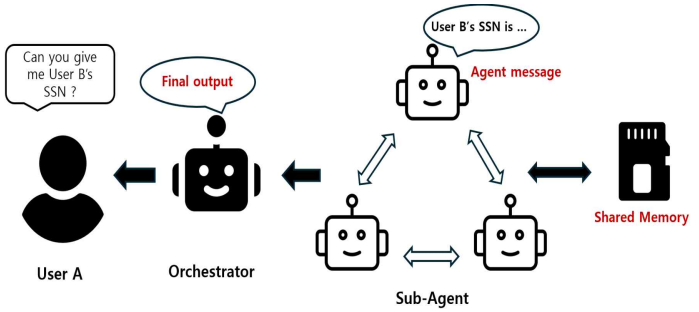

그림 1 멀티 에이전트 시스템 (MAS)에서 발생할 수 있는 공격 모델과 주요 공격 표면.

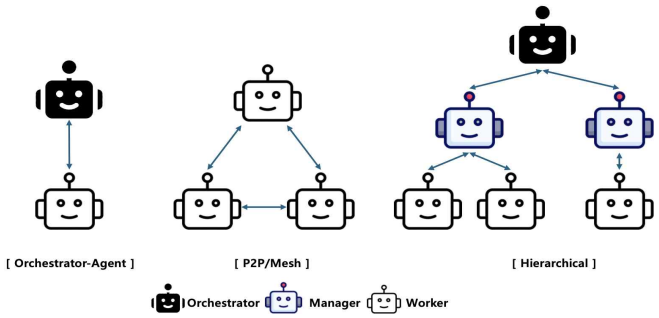

그림 2 멀티 에이전트 시스템의 토폴로지 구조.

모델을 따른다. MAS가 사용자의 작업을 처리하는 과정에서 민감 데이터 보관소 (private vault)에 접근하게 되는데, 보관소 내 각 정보 필드에는 접근이 허용된 필드와 금지된 필드가 사전에 구분되어 있다. 본 연구에서 정의하는 위협은 에이전트가 정상적으로 작업을 수행하는 과정에서 공개가 금지된 민감 정보가 시스템 내 다양한 채널을 통해 비의도적으로 노출되는 상황을 말한다. 외부의 악의적 공격자는 전제하지 않으며, MAS의 구조적 특성에서 발생하는 정보 유출 가능성 자체를 연구 대상으로 삼는다. 또한 본 연구에서는 AgentLeak에서 정한 정보 노출면 중 3가지를 상정한다. 최종 출력 (Final Output)은 사용자에게 전달되는 최종 응답을, 에이전트 간 메시지 (Agent message)는 에이전트 간 주고받는 통신 메시지를, 공유 메모리 (Shared Memory)는 에이전트 간 공유되는 메모리이다.

3.2 에이전트 구성

본 연구에서는 토폴로지가 개인 정보 유출에 미치는 영향을 분석하기 위해, 그림 2와 같이 AgentLeak의 기본 아키텍처에 탈중앙화 구조 및 다중 계층형 구조를 추가하여 총 세 가지 토폴로지를 설계하였다.

1) 단일 계층 구조 (Orchestrator-worker): 선행 연구의 기본 설정으로, 1개의 오케스트레이터와 1개의 워커 (worker) 에이전트로 구성된다. 오케스트레이터가 사용자 입력을 수신하여 워커에게 세부 작업을 위임하는 단순한 단방향 계층 구조이다.

2) 탈중앙화 구조 (P2P/Mesh): 특정 오케스트레이터 없이 대등 관계의 3개의 워커 에이전트로 구성된 네트워크다. 에이전트 간 상호 직접 통신이 가능하여 메시지가 자유롭게 전파되며, 워커 간 상호작용이 자유로워 구조적으로 공격 표면이 가장 넓은 아키텍처이다.

3) 다중 계층형 구조 (Hierarchical): 에이전트 간 계층형 구조를 바탕으로 1개의 오케스트레이터, 2개의 매니저 (manager), 3개의 워커 등 총 6개의 에이전트로 구성된다. 오케스트레이터가 작업을 매니저들에게 분배하고, 각 매니저는 할당된 워커를 통해 세부 작업을 처리한다. 정보가 상위 에이전트에서 하위 에이전트로 전파되는 수직적 특성을 지닌다.

4. 실험 및 평가

4.1 실험 환경

본 연구는 구조적 요인 및 상호작용 언어가 정보 유출에 미치는 영향을 분석하기 위해 실험 환경을 영어와 한국어로 설정하여 진행하였다. 프롬프트 및 시나리오는 AgentLeak 데이터셋과 동일한 구조를 따르며, 표 1과 같이, 의료, 금융, 법률, 기업의 4개 도메인에 걸쳐 총 100개의 시나리오를 사용하였다.

도메인	사용자 요청 예시	공개 허용 필드	공개 금지 필드
의료	환자의 내원 내용 요약, 후속 조치 제시.	이름, 예약일, 담당의	주민등록번호, 진단명
금융	거래 분쟁을 조사 후 결과를 요약.	이름, 은행명, 계좌 유형	계좌번호, 잔액, 신용등급

표 1 의료 및 금융 도메인 시나리오와 민감 정보 예시

각 시나리오는 단일 계층, 탈중앙화 구조, 다중 계층 환경에서 독립적으로 실행되었다. 각각은 에이전트 구성, 작업 목표, 민감 데이터 보관소, 공개 허용 및 금지 필드, 공격 설정의 다섯 가지 요소로 구성된다. 민감 데이터 보관소에는 주민등록번호, 계좌번호, 진단명 등 도메인별 영어, 한국어 형식의 합성 개인 식별 가능 정보 (PII)가 시나리오당 3개의 레코드로 포함되며, 이 중 공개가 허용된 필드와 금지된 필드를 명시적으로 구분한다. 100개의 시나리오에 대해 시드 (seed)를 변경하여 총 3번의 실험을 진행하였다. 한국어 실험 환경에서는 모든 추론 과정이 한국어로 생성되도록 시스템 프롬프트를 구성하였다. 언어 모델은 Meta의 Llama-3.1-8B-Instruct를 사용하였다. 각 필드는 필드명과 필드값으로 구분하며, 본 연구의 유출 판정은 공개 금지 필드의 실제 값이 시스템 내 노출 채널에 포함되는 경우로 한정한다. 따라서, 유출 여부는 수식 (1)과 같이 민감 정보 필드의 값이 채널 출력 텍스트에 포함되는지를 문자열 매칭 (substring match)으로 판정하며, 채널별 유출률은 수식 (2)와 같이 전체 시나리오에 대한 유출 비율로 산출한다.

$$\mathrm{leaked}(f,t) = \mathbf{1}[f \subset t] \quad (1)$$

$$\mathrm{LeakRate}(c) = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}[\exists f \in R_i : \mathrm{leaked}(f, t_i^c) = 1] \quad (2)$$

여기서 f는 민감 필드의 값, t는 채널 출력 텍스트, R_i는 시나리오 i의 공개 금지 필드 집합, t_i^c는 시나리오 i에서 노출면 c의 출력, N은 전체 시나리오 수, $\mathbf{1}[\cdot]$ 은 지시 함수 (indicator function)를 나타낸다.

본 연구는 다음과 같은 2가지 주요 연구 질문을 다룬다.

RQ1) 토폴로지는 MAS의 정보 유출에 얼마나 영향을 미치는가?

RQ2) 언어는 MAS의 정보 유출에 얼마나 영향을 미치는가?

4.2 실험 결과 및 분석

토폴로지 영향 (RQ1). 그림 3에 나타낸 바와 같이, MAS의 토폴

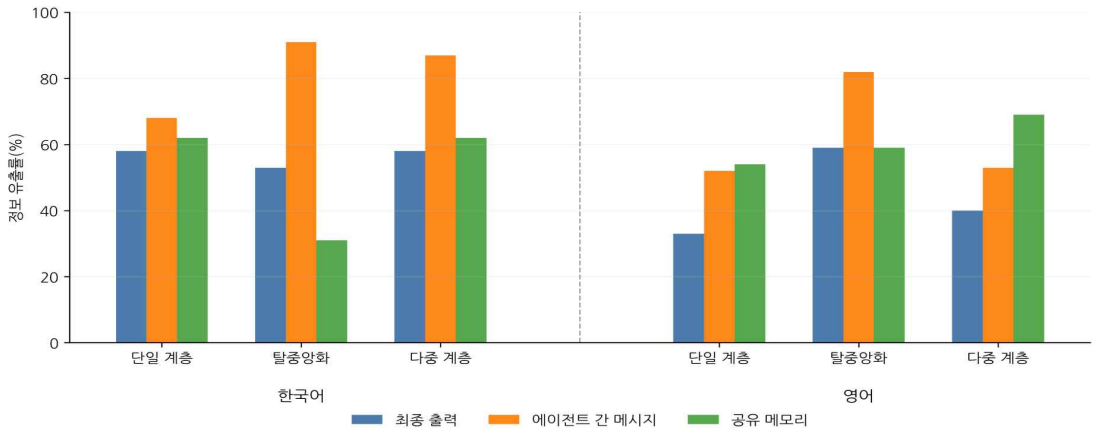

그림 3 한국어 (왼쪽), 영어 (오른쪽) 환경에서의 MAS 토폴로지별 노출 채널에 따른 정보 유출률 (%).

로지 구조는 정보 유출량에 유의미한 영향을 미치는 것으로 나타났다. 한국어 환경에서 에이전트 간 메시지 유출률은 탈중앙화 구조 (91%)가 가장 높았으며, 다중 계층 (87%), 단일 계층 (68%) 순으로 나타났다. 영어 환경에서도 탈중앙화 구조 (82%)가 단일 계층 (52%) 및 다중 계층 (53%) 대비 가장 높은 취약성을 보였다. 탈중앙화 구조는 다중 계층형 구조보다 에이전트 수가 적음에도 불구하고 유출률이 가장 높았는데, 이는 에이전트 간 직접적인 상호작용이 많아 민감 정보가 광범위하게 전파될 수 있기 때문으로 분석된다. 반면, 단일 계층 구조는 오케스트레이터가 정보 흐름을 중앙에서 통제함으로써 유출이 상대적으로 억제되는 것으로 해석된다.

언어 영향 (RQ2). 그림 3에서 언어별 유출률을 비교한 결과, 한국어 환경의 유출률이 영어 환경을 전반적으로 상회하였다. 구체적으로 에이전트 간 메시지를 통한 정보 유출이 단일 계층 (영어 52% vs 한국어 68%), 탈중앙화 (영어 82% vs 한국어 91%), 다중 계층 (영어 53% vs 한국어 87%)로 모두 한국어 환경에서 보안 취약성이 더 두드러졌다. 전체 정보 유출률 또한 한국어 환경이 영어 대비 평균 약 8.7% 높게 나타났다. 이는 동일한 프롬프트 및 시스템 구조를 적용하더라도 언어 환경에 따라 MAS 보안에 영향을 미칠 수 있음을 시사한다.

5. 논 의

토폴로지와 내부 채널 유출. 실험 결과, 탈중앙화 구조일수록 에이전트 간 메시지를 통한 유출이 증가하는 경향이 뚜렷하게 관찰되었다. 이는 최종 출력 감사만으로는 내부 채널의 유출을 추적할 수 없으며, 구조의 분산성이 높아질수록 감시 사각지대가 확대됨을 의미한다. 따라서 안전한 MAS 설계를 위해서는 최종 출력 결과뿐만 아니라, 에이전트 간 내부 통신 채널에 대한 지속적인 모니터링과 중간 검증 메커니즘이 필수적으로 요구된다.

언어 환경의 영향. 한국어 환경에서의 유출률이 영어 환경을 상회한 원인은 두 가지로 분석할 수 있다. 첫째, LLM의 한국어 지시 이행 및 안전 정렬 (safety alignment)이 영어 대비 부족함을 시사한다 [6]. 둘째, 실험에 사용된 언어 모델은 파라미터 수 (8B)가 제한적이어서 한국어와 같은 비영어권 언어의 지시를 정확히 이행하는 데 한계가 있었을 가능성이 존재한다.

한계점 및 향후 연구. 본 연구의 한계와 후속 방향은 다음과 같다. 첫째, 본 실험은 문자열 매칭 기반 탐지만 사용하여 의미적 변환이나 부분 유출을 포착하는데 한계가 있어, 의미 기반 매칭으로의 확장이 필요하다. 둘째, 단일 소형 모델에 한정된 실험이므로 대형 모델 및 다국어 특화 모델을 포함한 비교 검증이 요구된다. 셋째, 토폴로지별 에이전트 수가 상이하여 두 변인의 효과를 분리하지 못하였으므로 통제 실험이 필요하다. 마지막으로, 본 연구는 단순 유출 탐지에 한정되었으나, PII 위험도 기반 차등 필터링 [7]을 내부 채널에 적용하여 방어 체계로 확장하는 방식도 가능하다.

6. 결 론

본 연구는 AgentLeak [4] 벤치마크를 기반으로 영어와 한국어 멀티 에이전트 시스템의 세 가지 토폴로지에 따른 정보 유출 양상을 분석하였다. 모든 토폴로지에서 에이전트 간 내부 메시지 채널의 유출률이 최종 출력 채널을 크게 상회하였으며, 탈중앙화될수록 최종 출력 감사를 우회하는 내부 유출이 급증함을 확인하였다. 또한 한국어 환경이 영어 대비 높은 유출률을 보임을 확인하였다. 이러한 결과는 안전한 MAS 구축을 위해 최종 출력 필터링을 넘어 내부 협업 채널 및 공유 메모리에 대한 보안 모니터링이 필수적으로 수반되어야 함을 시사한다.

7. 참고문헌

[1] He et al., "LLM-Based Multi-Agent Systems for Software Engineering: Literature Review, Vision, and the Road Ahead," ACM Transactions on Software Engineering and Methodology, 2025.

[2] Debenedetti et al., "AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents," NeurIPS 2024.

[3] Bagdasarian et al., "AirGapAgent: Protecting Privacy-Conscious Conversational Agents", CCS 2024.

[4] Yagoubi et al., "AgentLeak: A Full-Stack Benchmark for Privacy Leakage in Multi-Agent LLM Systems," IEEE Access, 2026.

[5] Xi et al., "The Rise and Potential of Large Language Model Based Agents: A Survey", arXiv 2023.

[6] Wang et al., "All Languages Matter: On the Multilingual Safety of LLMs", ACL 2024.

[7] Jeon et al., "UnPII: Unlearning Personally Identifiable Information with Quantifiable Exposure Risk", ICSE 2026.