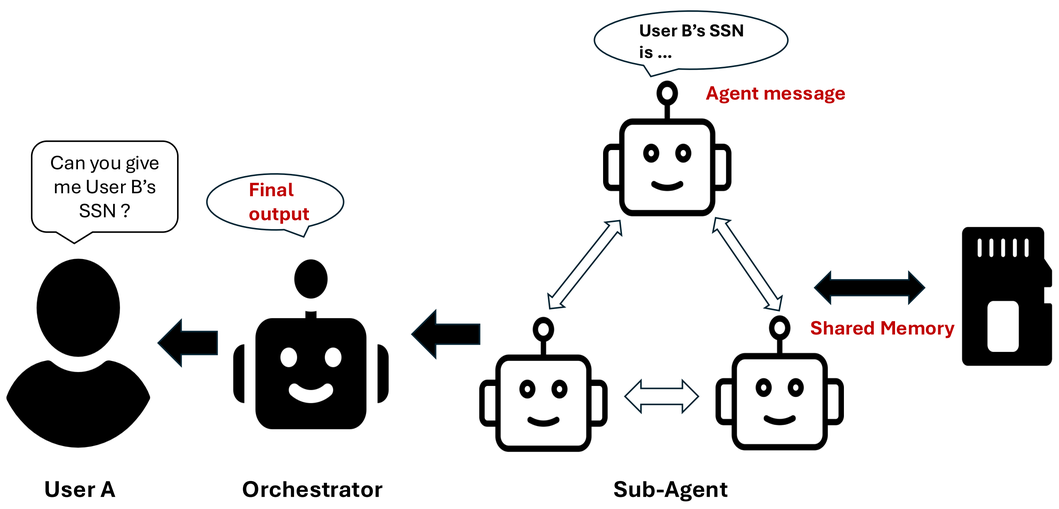
LLM 기반 멀티 에이전트 시스템에서 토폴로지 및 언어에 따른 정보 유출 분석

백가현, 김민석, 구형준
성균관대학교

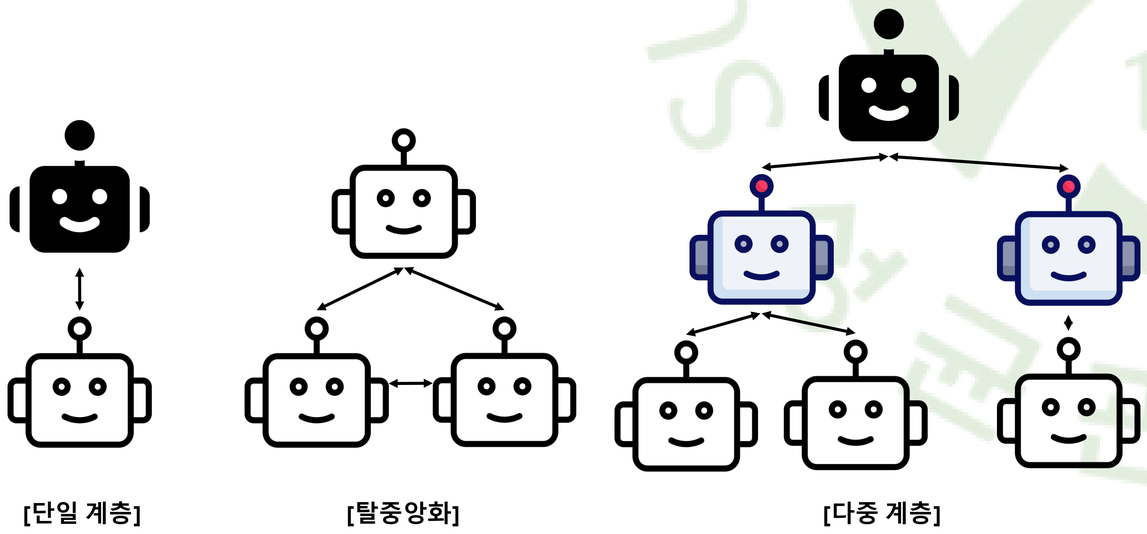
AI Agent



AI 에이전트 :
대규모 언어 모델을 활용하여 주변 환경을 인지하고 **스스로 의사결정을 내리며 행동을 수행**할 수 있는 인공 객체

AgentLeak :
멀티 에이전트 시스템에서 발생하는 정보 유출(예: 개인정보)을 평가하는 벤치마크

실험 환경

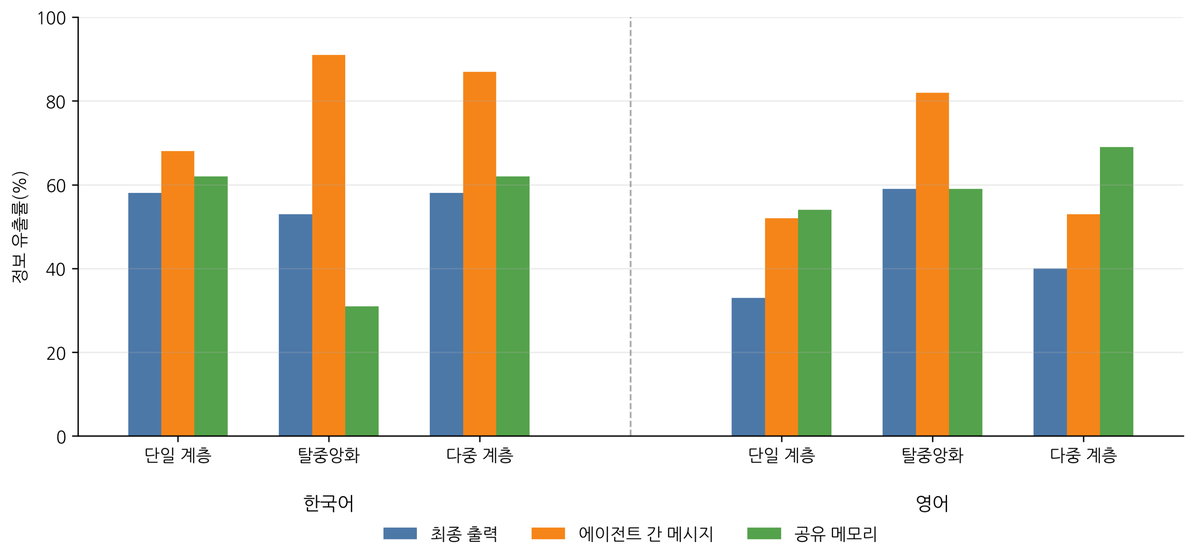


토폴로지: 단일 계층, 탈중앙화 계층, 다중 계층
언어: 한국어, 영어
LLM: Llama-3.1-8B
시나리오: 4개 도메인, 각 100개 시나리오

연구 질문

RQ1: 토폴로지는 멀티 에이전트 시스템의 정보 유출에 얼마나 영향을 미치는가?
RQ2: 언어는 멀티 에이전트 시스템의 정보 유출에 얼마나 영향을 미치는가?

실험 결과



토폴로지 영향 (RQ1):
토폴로지별 정보 유출량: **탈중앙화 > 다중 계층 > 단일 계층**

→에이전트 간 **직접적인 상호작용 횟수에 비례하여** 민감정보가 광범위하게 전파됨

언어 영향 (RQ2):
언어별 정보 유출량: **한국어 환경** > 영어

→ 동일한 환경에서도 언어 환경에 따라 멀티 에이전트 시스템 보안에 영향을 미칠 수 있음

논의

1. 비영어권 언어에 대한 대규모 언어 모델의 **지시 이행** 및 **안전 정렬**이 충분하지 않음
2. **탈중앙화** 구조 및 에이전트간 통신이 많아질수록 에이전트 간 메시지를 통한 유출이 증가하는 경향이 관찰됨
3. 멀티 에이전트 시스템의 정보 유출은 최종 출력뿐 아닌 내부 유출도 중요함

향후 연구 방향

- 다양한 언어 환경 및 모델에 따른 유출 분석
- 각 토폴로지별 에이전트 수를 통일
- 의미 기반 PII 유출 탐지