JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 11, pp. 379-386, November 2026
https://doi.org/10.9708/jksci.2026.31.11.379

Quantifying Internal Data Exposure in an Open-Source Multi-Tool LLM Agent for IT Service Management

You-Min Kim*, Jae-Woong Kim**, Won-Young Jeon***

*Student, IT Convergence Major, Kongju National University, Cheonan, Korea
**Professor, Dept. of Software, Kongju National University, Cheonan, Korea
***Lecturer, Dept. of Software, Kongju National University, Cheonan, Korea

[Abstract]

To address fragmented IT service management, this study implements a ReAct agent integrating structured asset queries, unstructured document retrieval, network diagnosis, and remote log collection using open-source frameworks and an open-weight model via a commercial inference API, and quantifies structural barriers to enterprise deployment. Across 30 trials, the invoked tools matched the expected tools in every query type except document retrieval (2 of 8), as the model answered from parametric knowledge instead of invoking the tool. Because the measured loop re-sends accumulated messages at each step, transmission amplification was 3.42 for structured queries and 3.80 for chains, mainly from re-sent results and errors of repeated asset queries. As a potential exposure at the retrieval stage, account-information chunks reached the top three for 2 of 20 general and 6 of 10 targeted queries, but none appeared in the four retrieval calls during agent runs. Transmission thus depends on how the loop handles messages, and prompt-level constraints ensured neither tool use nor compliance.

▸ **Key words**: LLM Agent, ReAct, Retrieval-Augmented Generation, Data Exposure, IT Service Management

[요 약]

본 연구는 분절된 IT 서비스 관리 문제를 해결하고자 오픈소스 프레임워크와 상용 API로 제공되는 가중치 공개 모델로 정형 자산 조회, 비정형 문서 검색, 네트워크 진단, 원격 로그 수집을 통합한 ReAct 에이전트를 구현하고, 기업 환경 배포의 구조적 제약을 정량 분석한다. 30회 시행에서 호출된 도구는 문서 검색(8건 중 2건)을 제외한 모든 유형에서 기대와 일치하였다. 측정한 실행 루프는 추론 단계마다 누적 메시지를 외부 API로 다시 보내므로 재전송 배수는 정형 조회 3.42, 다중 도구 연쇄 3.80이었고 반복된 자산 조회의 결과와 오류 메시지가 주된 요인이었다. 계정 정보 청크는 검색 계층에서 일반 질의 20건 중 2건, 계정 정보 요구 질의 10건 중 6건의 상위 3건에 들어 잠재적 노출을 보였으나 에이전트 실행의 문서 검색 4회에서는 반환되지 않았다. 전송량은 메시지 관리 방식에 좌우되며, 프롬프트 수준의 제약은 도구 호출과 지시 준수를 보장하지 못하였다.

▸ **주제어**: 거대언어모델 에이전트, 추론-행동 결합, 검색 증강 생성, 데이터 노출, 정보기술 서비스 관리

• First Author: You-Min Kim, Corresponding Author: Jae-Woong Kim
*You-Min Kim (dbals1222@naver.com), IT Convergence Major, Kongju National University
**Jae-Woong Kim (jykim@kongju.ac.kr), Dept. of Software, Kongju National University
***Won-Young Jeon (jun1young@kongju.ac.kr), Dept. of Software, Kongju National University
• Received: 2026. 09. 21, Revised: 2026. 11. 02, Accepted: 2026. 11. 07.

Copyright © 2026 The Korea Society of Computer and Information
http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945
www.kci.go.kr

I. Introduction

제조 기업의 IT 인프라는 시간이 지날수록 이질적으로 축적된다. 도입 시점이 다른 서버가 세대별로 혼재하고, 운영체제 버전이 갈라지며, 각 시스템의 운영 지침은 담당자와 시기에 따라 서로 다른 형식으로 남는다. 문제는 장애가 발생하는 순간에 드러난다. 관리자가 하나의 장애를 진단하려면 최소한 세 가지 정보가 필요하다. 대상 자산이 무엇이며 어떤 사양과 주소를 갖는가, 유사한 장애가 과거에 있었다면 어떤 절차로 해결하였는가, 현재 그 시스템이 어떤 상태에 있는가이다. 그런데 이 세 가지는 각각 자산 관리 대장, 문서 관리 시스템, 서버 콘솔이라는 서로 다른 위치에 존재한다. 관리자는 진단이 아니라 정보 취합에 시간을 소진하며, 이 과정은 담당자의 숙련도에 크게 의존한다.

대규모 언어 모델의 등장은 이 문제에 대한 접근 방식을 바꾸었다. 자연어 질의를 데이터베이스 질의문으로 변환하는 기법은 정형 데이터에 대한 접근 장벽을 낮추었고[1], 검색 증강 생성은 비정형 문서에서 질의와 의미적으로 유사한 내용을 찾아내어 어휘 불일치 문제를 완화하였다[2]. 나아가 ReAct 계열의 도구 사용 에이전트는 언어 모델이 스스로 필요한 도구를 선택하고 그 결과를 관찰하여 다음 행동을 결정하는 구조를 제시하였다[3]. 그러나 이러한 기법을 실제 사내 인프라 운영에 적용하려 할 때 제약이 나타난다. 상용 통합 플랫폼은 도입은 빠르나 공급자의 정책 변경과 서비스 종료에 그대로 노출된다. 조회만 가능한 시스템은 무엇이 존재하는가에는 답하지만 현재 어떤 상태인가에는 답하지 못한다. 그리고 서버 호스트명, 내부 IP 주소, 계정 정보, 시스템 로그와 같이 외부로 반출되어서는 안 되는 정보가 추론을 위해 외부 API로 전송될 수 있다.

기존의 적용 연구는 대체로 구현 가능성의 입증에 초점을 두고 종료된다. 본 연구는 구현에 더하여, 그 구현을 운영 환경에 그대로 적용하기 어려운 이유를 실행 구조 수준에서 규명하고 이를 정량 측정한다는 점에서 차별성을 갖는다. 구체적으로 다음 세 가지 연구 질문을 설정하였다. 첫째, 오픈 생태계의 구성요소만으로 정형·비정형 데이터 조회와 대상 시스템의 상태 진단을 통합한 에이전트를 구성할 수 있으며, 그 도구 선택은 질의 유형에 관계없이 신뢰할 수 있는가. 둘째, 상용 API를 추론 엔진으로 사용하는 현행 구조에서 사내 인프라 정보가 외부로 전송되는 규모는 어느 정도이며, 그 규모를 결정하는 요인은 무엇인가. 셋째, 도구 명세에 자연어로 기술된 프롬프트 수준의 통제는 실효적인가. 다만 온프레미스 전환 등 여러 환경에 따른 응답 품질 비교는 본 논문의 범위를 벗어나며, 후속 연구의 과제로 한다.

II. Related Works

2.1. Information Fragmentation and Retrieval

IT 서비스 관리 체계는 자산 관리, 변경 관리, 장애 관리를 별개의 프로세스로 정의한다. 이러한 분리는 프로세스 관점에서는 타당하나 데이터 관점에서는 파편화를 낳는다. 자산 정보, 장애 이력, 절차 지식이 서로 다른 시스템에 남기 때문이다. 이 중 정형 자산 정보에는 자연어 질의를 SQL로 변환하는 Text-to-SQL을 적용할 수 있으며, 대표적 평가는 질의와 함께 데이터베이스 스키마를 입력으로 제공한다[1]. 본 구현은 컬럼 구성을 제공하지 않았다(3.3절).

기존 지식 관리 시스템은 이 파편화를 검색으로 해소하고자 하였으나, 대다수는 어휘 빈도에 기반한 검색 모형에 의존한다[4]. 어휘 기반 검색은 질의어와 문서의 단어가 일치할 때에만 작동하므로, 실무자가 사용하는 표현과 문서에 기록된 표현 사이의 괴리를 넘지 못한다.

검색 증강 생성은 언어 모델이 답변을 생성하기 전에 외부 지식원에서 관련 문서를 검색하여 문맥으로 제공하는 기법이다[2]. 문서와 질의를 동일한 임베딩 공간의 벡터로 변환하고 벡터 간 거리로 유사도를 판정하므로, 표층 어휘가 달라도 의미가 가까우면 검색된다[5]. 검색 성능은 임베딩 모델의 언어 적합성에 좌우되며, 문장 단위 의미 표현을 학습한 모델이 문서 검색에 유리하다[6]. 국내에서도 RAG의 아키텍처와 최적화 동향에 대한 검토가 이루어지고 있다[7].

2.2. Tool-Using Agents and Data Exposure

ReAct는 언어 모델이 추론과 행동을 교대로 수행하도록 하는 기법이다[3]. 모델은 현재 상황에 대한 사고를 생성하고, 그에 따라 도구를 호출하며, 도구의 반환값을 관찰한 뒤 다시 사고를 생성한다. 단계적 사고의 명시가 추론 성능을 향상시킨다는 관찰[8]이 이 구조의 이론적 기반을 이루며, 국내에서도 에이전트 구조에 대한 검토가 진행되고 있다[9].

ReAct의 핵심은 도구 반환값이 후속 추론의 입력이 된다는 점이다. 에이전트 프레임워크는 도구 실행 결과를 도구 메시지로 대화 목록에 덧붙여 다시 모델에 전달하며, 무엇을 다시 전달할지는 기법이 아니라 프레임워크의 메

www.kci.go.kr

시지 관리 방식이 정한다. 이 경로가 본 연구가 측정하는 전송 경로이다.

도구 사용 에이전트의 데이터 유출은 최근 활발히 연구되고 있으며, 프롬프트 주입을 통한 유도, 메모리 추출, 다중 에이전트 간 전파 등이 주요 위협으로 분류된다[10,11]. 이들 연구는 공격자가 존재하는 적대적 상황을 전제하거나, 도구 반환값에 악성 지시를 삽입하는 상황을 모사한 벤치마크 환경에서의 평가를 중심으로 한다[12]. 폐쇄망 환경을 위한 로컬 모델 적용 연구[13]가 유사한 문제의식을 공유하나, 전송량 자체를 측정 대상으로 삼지는 않는다.

반면 공격이 개입하지 않는 정상 동작 상태에서 실행 루프의 구조 자체로 인하여 발생하는 전송량, 특히 다단계 추론에서의 누적 재전송 규모를 측정한 사례는 상대적으로 드물다. 본 연구는 이 지점에 위치한다(Table 1).

Table 1. Related Work and This Study

Line of work	Prior focus	This study
ITSM [14]	Separate processes	Unified tool access
Text-to-SQL [1]	Accuracy	Column guessing
RAG [2,4-7]	Retrieval	Retrieved text sent
Agents [3,8,9]	Tool reasoning	Re-sent tool output
Security [10-13]	Attacks, local LLM	Normal-use sending

III. System Design

3.1. Design Requirements

본 시스템은 네 가지 요구사항을 전제로 설계하였다. 첫째, 추론 엔진을 교체하여도 도구 정의나 데이터 연동이 영향받지 않는 모델 독립성이다. 둘째, 사용자가 저장소를 구분하여 질의하지 않아도 질의 성격에 따른 경로 선택이 시스템 내부에서 이루어지는 이기종 데이터 통합 접근이다. 셋째, 기록된 정보의 조회를 넘어 대상 시스템의 현재 상태를 직접 관측하는 수단이다. 넷째, 상용 통합 플랫폼에 의존하지 않고 구성요소를 개별적으로 교체할 수 있는 오픈 생태계 기반 구성이다. 이를 계층 분리(3.2절), 명세 기반 판단(3.3절), 진단·로그 도구(3.4절), 교체 가능한 구성요소(Table 3)로 구현하였다. 도구 선택과 인자는 코드로 검증하지 않는다(5.2절).

3.2. System Architecture

시스템은 사용자 질의를 입력받는 인터페이스, 도구 선택과 추론을 담당하는 에이전트 계층, 실제 데이터에 접근하는 도구 계층, 그리고 정형·비정형 저장소와 대상 서버로 구성되는 자원 계층으로 이루어진다. 전체 구조는 Fig. 1과 같다.

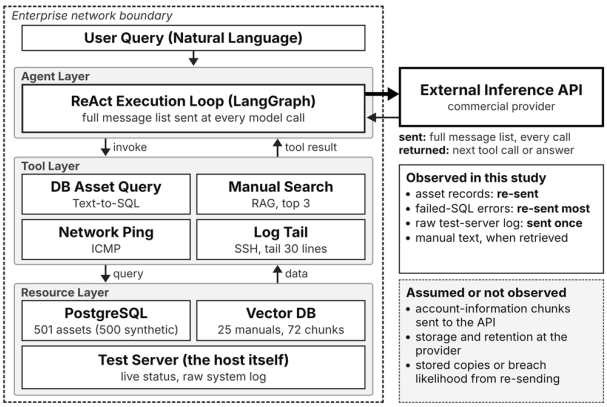


Fig. 1. System Architecture, Trust Boundary, and Observed versus Assumed Exposure

핵심은 에이전트 계층과 도구 계층의 분리이다. 에이전트 계층은 어떤 도구를 어떤 순서로 호출할지 결정할 뿐, 각 도구의 내부 구현을 알지 못한다. 도구 계층은 자연어 인자를 받아 실행 결과를 문자열로 반환한다. 이 분리가 모델 교체 시 도구 정의를 수정하지 않아도 되는 성질을 보장한다.

3.3. Heterogeneous Data Layer

정형 데이터 접근은 자연어 질의를 SQL로 변환하는 방식을 사용한다. 도구 명세에는 테이블명이 제공되나 컬럼 구성은 제공되지 않으며, 모델은 컬럼명을 추정하여 질의문을 생성한다. 생성된 질의문은 실행 전에 마크다운 코드 펜스를 제거하는 전처리를 거친다. 비정형 데이터 접근은 검색 증강 생성 방식을 사용한다. 사내 운영 문서를 마크다운 제목 단위로 나눈 뒤 500자 단위로 분할하고 50자를 중첩시켜 색인하였으며, 임베딩 모델은 한국어 문장 임베딩에 특화된 모델을 사용하였다. 검색은 상위 3건을 반환하도록 설정하였다.

두 경로 가운데 어느 것을 사용할지는 사전 규칙이 아니라 도구 명세에 대한 모델의 판단으로 결정된다. 질의 유형과 도구를 대응시키는 조건문은 코드에 존재하지 않는다.

3.4. Agent Execution Model and Tools

에이전트는 추론과 행동을 교대로 수행하는 ReAct 방식으로 동작한다. 모델이 현재 메시지 목록을 입력받아 다음 행동을 결정하고, 도구가 호출되면 그 반환값이 메시지 목록에 도구 메시지로 추가되며, 갱신된 목록 전체가 다시

www.kci.go.kr

모델에 전달된다. 이 순환은 모델이 도구 호출을 더 이상 요구하지 않을 때까지 반복된다.

여기서 주목할 성질은 메시지 목록이 매 단계 누적된 상태로 전송된다는 점이다. 본 구현은 메시지를 요약하거나 걸러 내지 않으므로 초기 단계에서 반환된 도구 출력은 이후 모든 단계의 입력에 포함된다. 이 성질은 본 구현에서 다단계 추론의 연속성을 유지하는 방식인 동시에, 제Ⅴ장 1절에서 분석할 데이터 전송 문제의 직접적 원인이 된다. 본 시스템은 네 개의 도구를 구현하였으며, 명세는 Table 2와 같다.

Table 2. Tool Specification

Tool	Input	Action	Return
DB_Asset_Query_Tool	SQL (model-generated)	Code-fence removal, SQL execution (no filtering)	Query result
Manual_Search_Tool	Natural language	Vector search (k = 3)	Retrieved chunks
Network_Ping_Tool	IP address	ICMP echo, 4 packets, 10 s timeout	Standard output
Log_Tail_Tool	IP address	SSH, sudo tail of the last 30 log lines	Stdout or stderr

원격 접속 도구의 자격증명은 환경변수나 실행 시 입력으로 메모리에만 두어 메시지에 포함되지 않는다. 원격 명령은 로그 마지막 30줄을 읽는 tail 하나이며, 비밀번호 없는 권한 상승도 이 명령으로 한정하였다. 오류 원인 구분을 위해 표준 출력과 표준 오류를 함께 읽는다.

도구 사용에 대한 제약은 도구 명세에 자연어로 기술되어 있다. 반환 행 수를 다섯 건으로 제한할 것, 테이블명에 인용 부호를 사용할 것, 원격 로그 조회는 네트워크 진단이 정상일 때에만 수행할 것이 여기에 해당한다. 이 제약들은 프롬프트 수준의 지시이며 코드가 검사하지 않는다. 질의문의 종류도 제한하지 않으며, 데이터베이스 계정은 쓰기 권한을 가졌다(기록이 남은 실행에서 데이터를 바꾸는 질의문은 실행되지 않았다).

3.5. Experimental Setup

구현은 Ubuntu 환경의 Python 3.10 가상환경에서 수행하였으며, 실행 구성과 데이터는 Table 3과 같다. Apache 2.0 라이선스로 가중치가 공개된 모델을 상용 API(Groq, 무료 요금제)로 호출하였으므로 모델에 전달되는 메시지는 모두 외부로 전송된다.

Table 3. Experimental Dataset and Configuration

Category	Description
Structured	501 server assets (Ubuntu 22.04: 141, Rocky Linux 8: 126, Windows Server 2019: 122, Windows Server 2016: 111, reachable test server: 1); 500 synthetic, OS randomly assigned
Unstructured	25 internal IT operation manuals
Chunking	500 characters, 50-character overlap, 72 chunks
Retrieval	ko-sroberta-multitask, Chroma 1.5.9, Top 3
Model	openai/gpt-oss-20b, Groq API, temp. 0
Agent	LangGraph 1.2.11, no system prompt/memory
Loop	Default limit 10,007; stops at no tool call

합성 데이터는 실제 환경의 규모와 운영체제 종류를 참고하였다. 원격 진단 대상은 실험을 수행한 개발 서버 자신(루프백)이므로 별도 장비의 진단을 실증한 것은 아니며, 로그 원문은 싣지 않는다.

Ⅳ. Experimental Results

4.1. Evaluation Design

도구 선택은 연쇄 외 질의에서 호출된 도구 집합이 기대 집합과 같으면 일치로 판정하였다(반복 허용, 기대 외 도구는 불일치). 반복·불필요 호출과 답변 정확성은 따로 집계하였다. 평가 질의는 다섯 유형 26건(Table 4)이며, 연쇄 질의 2건은 각 3회 반복하여 총 30회를 시행하였다. 매뉴얼 검색 질의는 색인 문서의 어휘를 직접 사용하지 않고 작성하여, 문자열 일치가 아닌 의미 기반 검색이 요구되도록 하였다.

다중 도구 연쇄 질의는 조건부로 채점하였다. 원격 로그 조회는 네트워크 진단이 정상일 때에만 수행하도록 명세되어 있으므로, 진단 성공 시 호출을, 실패 시 미호출을 각각 정답으로 판정하였다. 자산 조회가 진단보다 먼저여야 하며 진단 주소가 대상인지는 따로 판정하였다. 다만 본 실험의 연쇄 대상 서버는 항상 도달 가능하였으므로, 진단 실패 조건은 관측되지 않았다. 따라서 실패 시 로그 미호출 분기는 검증되지 않았다(제Ⅴ장 3절).

측정은 프레임워크의 콜백 인터페이스를 통해 수행하였으며, 시스템 구현 코드는 수정하지 않았다. 도구 출력은 반환 문자열 길이의 합, 전송량은 모델 호출마다 요청에 든 도구 메시지 길이의 누적 합이며 사용자 질의·모델 응답·도구 인자와 정의는 제외하였다. 지연은 질의 시작부터 최종 답변까지의 시간으로 API 응답, 도구 실행, 사용량 한도 대기와 재시도를 포함하며, 신뢰구간은 Wilson 95% 구간이다.

www.kci.go.kr

도구 호출 횟수는 도구가 실제로 실행된 횟수를, 모델 호출 횟수는 외부 추론 API에 요청이 전송된 횟수를 의미한다. 한 번의 모델 호출에서 복수의 도구 호출이 요청될 수 있으므로 두 값은 일치하지 않으며, 누적 메시지의 재전송 횟수를 결정하는 것은 모델 호출 횟수이다. 사용량 한도 오류로 다시 실행한 2회는 완료된 시도의 값으로 집계하였고(중단된 시도의 전송 1,217자, 9,471자), 지연만 두 시도의 합이다.

4.2. Tool Selection

측정 결과는 Table 4와 같다. 매뉴얼 검색을 제외한 모든 유형에서 기대 도구와 일치하였고, 연쇄는 모두 자산 조회를 먼저 수행하였으나 1회는 대상 특정 전에 진단하였다(4.3절). 불필요한 호출은 연쇄의 문서 검색 1회였다. 최종 답변은 네트워크·도구 불필요 8건이 적절하였고, 정형 조회 8건 중 3건은 다섯 건만 받은 결과를 전체처럼 제시하거나 조건을 잘못 적용하였다. 연쇄는 답변이 남은 질의 2의 3회 모두(질의 1은 기록 유실) 대상과 로그를 맞게 보고하였으나 2회는 서버명과 호스트명을 바꿔 적었고 1회는 로그에 없는 연도를 붙였다.

Table 4. Tool Selection by Query Type

Query type	N	Match (%)	95% CI	Calls (range)	Rep./ Extra
Structured query	8	8 (100.0)	67.6-100	2-12	37/0
Manual retrieval	8	2 (25.0)	7.1-59.1	0-1	0/0
Network diagnosis	4	4 (100.0)	51.0-100	1	0/0
Multi-tool chain	6	6 (100.0)	61.0-100	8-15	47/1
No tool required	4	4 (100.0)	51.0-100	0	0/0
Total	30	24 (80.0)	62.7-90.5	-	84/1

반면 매뉴얼 검색은 8건 가운데 2건에서만 도구가 호출되어 25.0%에 그쳤다. 코드 확정 전의 사전 실행에서도 같은 2건만 호출되었으나, 코드와 측정 도구가 다르고 당시 도구 명세를 확인할 수 없어 같은 조건의 재현으로 보지 않는다. 이 값은 8개 질의의 구성에 따른 것이며 일반적인 실패 확률로 해석할 수 없다(제Ⅴ장 2절).

4.3. Multi-Tool Chain Reasoning

연쇄 질의는 자산 조회로 대상 서버를 특정하고, 네트워크 진단으로 도달성을 확인한 뒤, 원격 접속으로 시스템 로그를 수집하는 3단계 구조이다.

6회 시행 전부에서 3단계 연쇄가 완주하였다. 수집된 로그에는 대상 호스트명, 프로세스 식별자, 사용자 식별자, 서비스명이 포함되어 있었으며, 이 값들은 후속 추론을 위해 외부 추론 API로 전송되었다.

한편 도구 호출 궤적은 시행마다 상당한 변동을 보였다. 도구 호출은 질의 1에서 8~9회, 질의 2에서 12~15회였고, 자산 조회 도구를 반복 호출한 뒤 진단으로 넘어가는 경우와 진단과 조회를 교대로 수행하는 경우가 모두 관측되었다. 질의 2의 한 시행은 대상 특정 전에 자산 목록 밖 주소로 진단하였다. 온도를 0으로 설정하였음에도 궤적의 결정성이 보장되지 않음을 보이는 결과이다.

4.4. Response Latency

측정 결과는 Table 5와 같다. 정형 조회는 최소 2.12초에서 최대 142.28초로 60배 이상 차이가 났다. 단순 집계 질의는 2회 호출로 완료된 반면, 조건에 부합하는 서버 목록을 요구하는 질의에서는 12회의 반복 조회가 발생하였다.

Table 5. Response Latency by Query Type (seconds)

Query type	N	Mean	Median	Min	Max
Structured query	8	39.59	32.26	2.12	142.28
Manual retrieval	8	4.90	4.58	1.72	12.77
Network diagnosis	4	6.44	4.86	4.58	11.48
Multi-tool chain	6	129.76	104.58	81.29	225.99

반복 조회의 주된 요인은 스키마 미제공에 따른 컬럼명 추정 실패와 재시도였다. 실행 기록이 남은 11회의 자산 조회 77회 중 32회가 실패하였고(없는 컬럼 28회, 테이블명 인용 누락 4회), 행 수 제한 지시는 컬럼 목록 조회까지 잘라 재조회를 낳았다. 매뉴얼 검색의 낮은 지연은 미호출 6건 때문이며 성능 우위가 아니다.

4.5. Data Transmission to the External API

측정 결과는 Table 6과 같다. 값은 시행당 평균 문자수(괄호는 범위)이며, Amp는 시행별 배수의 평균이므로 평균끼리 나눈 값(정형 4.13, 연쇄 3.96)과 다르다. 도구가 호출되지 않은 시행은 집계에서 제외하였다.

Table 6. Tool Output Sent to the External API

Query type	N	Tool output	Transmitted	Amp	LLM Calls
Structured query	8	1,769	7,304 (688-26,418)	3.42 (1.95-6.94)	6.6
Manual retrieval	2	1,198	1,198	1.00	2.0
Network diagnosis	4	340	340	1.00	2.0
Multi-tool chain	6	5,637	22,309 (15,739-46,224)	3.80 (3.02-5.98)	12.0

단일 도구만 호출되는 질의에서 배수는 1.00으로, 도구 출력이 한 번만 전송된다. 반면 도구가 반복 호출되는 정

www.kci.go.kr

형 데이터 조회에서는 3.42배, 다중 도구 연쇄에서는 3.80배로 증가하였으며, 시행별 최대 배수는 정형 조회 6.94배, 다중 도구 연쇄 5.98배였다.

다중 도구 연쇄의 평균 전송량은 22,309자이며, 단일 시행 최대값은 46,224자였다(도구 15회, 모델 16회 호출). 전송된 내용에는 서버 호스트명, IP 주소, 운영체제 버전, 자원 할당량(실험 서버 외에는 합성 값)과 함께 대상 서버의 시스템 로그 원문이 포함된다. 전송량의 75.8%(정형)와 74.7%(연쇄)는 재전송분이었고, 정형 조회 전송량의 80.6%는 실패한 조회의 오류 메시지였다. 실행 기록이 남은 연쇄 3회에서 사후에 정한 형식 기준으로 센 IP 주소 58개와 자산 식별자 74개는 각각 274회, 564회 전송된 반면 로그 90줄은 마지막 단계에서 한 번씩만 포함되었다.

4.6. Account-Information Exposure at the Retrieval Layer

문서 검색 도구는 유사도 상위 3건의 원문을 그대로 반환한다. 색인된 25건의 72개 청크 중 4개에 계정명과 비밀번호 형식의 계정 정보가 기재되어 있었고, 이 청크를 상위 3건에 하나 이상 포함한 질의 수는 Table 7과 같다. 이 측정은 검색 계층만 실행하여 결과가 사용자나 외부로 전달되지 않았다.

Table 7. Queries Retrieving Account-Information Chunks

Query type	N	Retrieved	Rate
General operational	20	2	10.0%
Account-targeted	10	6	60.0%

계정 정보와 무관한 통상적 운영 질의 20건 가운데 2건(10.0%, 무작위 기준 16.0%)에서 해당 청크가 검색 결과에 포함되었고, 그 분량은 20건의 검색 결과 24,489자 중 970자(4.0%)였다. 계정 정보를 명시적으로 요구하는 질의 10건에서는 6건에서 해당 청크가 검색되었다. 95% 신뢰구간이 넓어(2.8~30.1%, 31.3~83.2%) 일반적인 노출 확률로 볼 수 없다. 본 측정의 문서 검색 4회(중단된 시도의 1회 포함)는 해당 청크를 반환하지 않았으나 호출이 드물었던 결과이므로, Table 7은 검색 단계의 잠재적 노출로 한정한다.

V. Discussion

5.1. Re-sending in the Measured Execution Loop

제Ⅳ장 5절의 측정은 전송량이 도구가 반환한 양과 일치하지 않음을 보인다. 측정한 실행 루프가 매 추론 단계마다 누적 메시지 전체를 전송하므로, 초기 단계에서 반환된 도구 출력은 이후 모든 단계에서 반복 전송된다. 이 구현에서는 진단의 정밀도를 높이기 위해 도구 호출을 늘리는 선택이 전송량 증가로 직결된다.

다만 이는 측정한 메시지 관리 방식의 결과이며 ReAct 기법의 필연적 귀결로 일반화할 수 없다. 메시지 요약, 출력 필터링, 필드 선택은 재전송을 줄일 수 있으나 본 연구는 이를 비교하지 않았으므로, 이 결과를 다른 설계로도 해결되지 않는다는 근거로 볼 수 없다. 확인한 것은 행 수 제한 지시가 있었음에도 재조회와 재시도로 전송량이 늘었다는 점이다. 마스킹은 로그 같은 비정형 텍스트에서 민감 정보를 빠짐없이 식별하기 어려우며 본 연구에서 검증하지 않았다.

재전송은 노출되는 정보의 종류를 새로 늘리지는 않는다. 관측된 것은 자산 조회 결과와 오류 메시지가 여러 요청에 반복 포함된 것이다. 이것이 외부에 저장되는 사본이나 침해 가능성을 얼마나 늘리는지 측정하지 않았고 제공자 보존 정책도 조사하지 않았으므로, 요청 수만큼 처리·기록 기회가 늘 수 있다는 가능성을 위협 모델로만 제시한다(Fig. 1).

제Ⅳ장 6절의 결과는 여기에 검색 계층의 잠재적 노출 경로를 더한다. 유사도 기반 검색은 질의와 문서의 의미적 근접성만을 판단할 뿐, 요청자의 열람 권한을 판정하는 절차가 구조상 존재하지 않는다. 따라서 이 문제는 코퍼스 정제로 완화할 수는 있으나 제거할 수는 없으며, 문서 단위 접근 통제를 검색 계층에 도입하는 것이 근본적 대응이나 이는 별도의 설계 과제이다.

결론적으로 사내 인프라 정보의 외부 전송을 차단하면서 다중 도구 진단 기능을 유지하려면 추론 엔진 자체를 사내망에 위치시키는 방법이 가장 직접적인 해법으로 예측된다.

5.2. Limits of Prompt-Level Control

본 시스템의 도구 사용 제약은 모두 도구 명세, 즉 프롬프트 수준에 기술되어 있다. 코드는 이 제약들을 검사하지 않으며, 모델이 명세를 따를 것이라는 보장도 없다. 이 한계는 세 가지 형태로 관측되었다.

첫째, 도구 호출 자체가 보장되지 않는다. 근거 문서가 있는 매뉴얼 질의는 7건이었고(1건은 해당 매뉴얼이 없음) 그중 검색 도구가 호출된 것은 1건이었다. 더 중요한 문제는 미호출이 오류로 표면화되지 않는다는 점이다. 도구를 호출하지 않은 6건에 대하여 시스템은 모두 정상적으로 응답을 반환하였으나 그중 5건은 사내 실제 절차와 일치

www.kci.go.kr

하지 않았다. 모델은 사내에서 사용하는 시스템이 아닌 일반적으로 알려진 도구와 절차를 제시하였다.

특히 자산 조회 절차를 묻는 질의에는 도구 없이 SQL을 생성하였는데, 컬럼명 여섯 개는 스키마에 없었으나 테이블명 인용과 행 수 제한이라는 명세 형식은 따랐다. 명세가 프롬프트에 있다는 사실이 오히려 환각을 정교하게 만들 수 있다.

둘째, 명세가 준수될 때에도 부작용이 발생한다. 행 수 제한 지시는 컬럼 목록 조회에도 적용되어 재조회를 낳았고, 다섯 건만 받은 결과를 전체처럼 제시한 답변도 있었다. 셋째, 형식 지시도 지켜지지 않았다. 테이블명 인용 지시는 4회 어겨졌고, 조건부 지시는 대상 진단 실패가 없어 준수 여부를 확인하지 못하였다.

운영 환경 적용을 위해서는 프롬프트 수준의 통제를 코드 수준의 검증으로 전환하여야 한다. 구체적으로는 질의 성격에 따라 도구 호출을 강제하는 라우팅 계층, 도구 간 선행 조건을 코드로 검사하는 계층, 생성된 질의문을 실행 전에 구문 분석하여 조회 이외의 명령을 차단하는 검증 계층, 읽기 전용 데이터베이스 계정의 적용, 원격 명령의 허용 목록 관리가 필요하다.

5.3. Layered Dependency and Limitations

추론 모델이 제공자 측에서 중단되었을 때 대응은 모델 식별자 하나의 교체로 끝났으나, 에이전트 생성 함수는 프레임워크의 주 버전 갱신 이후 레거시 경로가 되었다. 지속가능성 위험은 모델 계층에만 있지 않으므로 구성요소의 판을 함께 기록하였다(Table 3).

본 연구에는 다음과 같은 한계가 있다. 첫째, 단일 조직의 환경에서 수행된 사례 연구이며 자산 규모와 문서 코퍼스가 제한적이고 자산은 합성 데이터이다. 둘째, 결과는 26개 질의 30회(연쇄 6회)의 구성에 따른다. 셋째, 검색 정확도에 대한 기준선 대비 평가를 수행하지 못하였다. 넷째, 단일 모델·제공자·실행 루프만 측정하였고, 전송 요인은 실행 기록 분해로 구분하였을 뿐 통제 실험으로 분리하지 않았으며, 클라이언트의 자동 재시도 요청은 전송량 계측에서 빠졌다. 다섯째, 원격 진단 대상이 실험 호스트 자신이고 대상 진단 실패가 없어 조건부 제약을 검증하지 못하였다.

다만 대화 목록 전체를 다시 보내는 방식으로 외부 추론 API를 호출하는 에이전트에는 같은 기제가 작동할 것이며, 그 규모는 메시지 관리 방식과 도구 출력 크기에 따라 달라진다.

VI. Conclusions

본 연구는 파편화된 IT 서비스 관리 환경의 정보 취합 병목을 해소하기 위하여, 오픈 생태계를 활용한 다중 도구 에이전트를 구현하고 그 구조적 제약을 정량 분석하였다.

첫째 연구 질문에 대하여, 오픈소스 프레임워크와 가중치 공개 모델(상용 API로 호출)로 네 도구를 통합한 에이전트를 구성하였으며, 3단계 연쇄 진단이 6회 시행 전부에서 완주함을 확인하였다. 그러나 도구 선택은 질의 유형에 관계없이 신뢰할 수 있는 수준은 아니었다. 문서 검색은 8건 중 2건에 그쳤으며, 도구를 호출하지 않은 응답이 오류로 표면화되지 않아 사용자가 답변의 출처를 구분할 수 없다는 문제가 확인되었다.

둘째 연구 질문에 대하여, 측정한 실행 루프가 매 추론 단계마다 누적 메시지를 재전송하므로 도구 출력이 외부 추론 API로 반복 전송됨을 확인하고 그 규모를 측정하였다. 재전송 배수는 정형 조회 3.42배, 다중 도구 연쇄 3.80배였으며, 다중 도구 연쇄의 평균 전송량은 22,309자였다. 실행 기록의 분해에서는 실패한 자산 조회의 재시도를 포함한 모델 호출 횟수가 주된 요인이었다. 계정 정보 청크는 검색 단계의 잠재적 노출로만 확인되었고 에이전트 실행에서는 반환되지 않았다.

셋째 연구 질문에 대하여, 프롬프트 수준의 통제는 실효적이지 않았다. 도구 호출 자체가 보장되지 않았고, 형식 지시가 어겨졌으며 행 수 제한 지시는 재조회를 낳았다. 조건부 지시는 실패 조건이 관측되지 않아 검증하지 못하였다.

이상의 결과는 현행 구조를 사내 인프라 운영에 바로 적용할 수 없음을 보인다. 도구 출력이 추론의 입력이 되는 한 외부 전송은 남으므로, 전송량을 줄이는 메시지 관리와 추론 위치의 선택이 필요하다.

향후 연구는 앞의 한계에 하나씩 대응한다. 다른 조직의 실제 자산 데이터와 더 큰 코퍼스에서 측정하고(첫째), 질의와 반복 시행을 늘리며(둘째), 검색 방식의 기준선 비교를 수행한다(셋째). 넷째 한계에 대해서는 호출 횟수와 출력 길이를 통제하고 클라이언트의 자동 재시도 요청까지 계측하여, 다른 모델·제공자와 메시지 요약·필터링을 적용한 루프를 비교한다. 본 연구에서 사용한 추론 모델은 가중치가 공개된 모델이므로 실제 운영 환경에 동일 모델을 배포할 수 있다. 이 경우 모델을 바꾸지 않고 추론 위치만 이전하게 되며, 서빙 환경의 차이에 따른 응답 품질 변화는 검증 과제로 남는다. 다섯째 한계에 대해서는 별도 장비와 진단 실패 조건에서 원격 진단과 조건부 제약을 시험한다. 아울러 생성된 질의문을 실행 전에 검증하는 계층의

www.kci.go.kr

도입, 도구 간 선행 조건의 코드 검사, 실행 시점에 자격증명을 발급하고 사용 후 폐기하는 접근 통제, 검색 계층의 문서 단위 권한 판정을 후속 과제로 한다.

REFERENCES

[1] T. Yu et al., "Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task," Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 3911-3921, Brussels, Belgium, Oct. 2018. DOI: 10.18653/v1/D18-1425

[2] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," Advances in Neural Information Processing Systems, Vol. 33, Online, Dec. 2020.

[3] S. Yao et al., "ReAct: Synergizing Reasoning and Acting in Language Models," Proceedings of the International Conference on Learning Representations, May 2023.

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

[5] V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 6769-6781, Online, Nov. 2020. DOI: 10.18653/v1/2020.emnlp-main.550

[6] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks," Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 3982-3992, Hong Kong, China, Nov. 2019. DOI: 10.18653/v1/D19-1410

[7] E. Lee and H. Bae, "A Survey on the Latest Research Trends in Retrieval-Augmented Generation," The Transactions of the Korea Information Processing Society, Vol. 13, No. 9, pp. 429-436, Sep. 2024. DOI: 10.3745/TKIPS.2024.13.9.429

[8] J. Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," Advances in Neural Information Processing Systems, Vol. 35, New Orleans, USA, Dec. 2022.

[9] S. Lim and J. R. Cho, "A Study on Multimodal LLM Agents," Journal of Information Technology and Applied Engineering, Vol. 15, No. 1, pp. 23-40, Apr. 2025. DOI: 10.22733/JITAE.2025.15.01.003

[10] F. He, T. Zhu, D. Ye, B. Liu, W. Zhou, and P. S. Yu, "The Emerged Security and Privacy of LLM Agent: A Survey with Case Studies," ACM Computing Surveys, Vol. 58, No. 6, pp. 1-36, Dec. 2025. DOI: 10.1145/3773080

[11] Y. Gan et al., "Navigating the Risks: A Survey of Security, Privacy, and Ethics Threats in LLM-Based Agents," arXiv, https://arxiv.org/abs/2411.09523

[12] E. Debenedetti, J. Zhang, M. Balunovic, L. Beurer-Kellner, M. Fischer, and F. Tramer, "AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents," Advances in Neural Information Processing Systems, Vol. 37, Vancouver, Canada, Dec. 2024.

[13] J. Heo, "A Local LLM-Based Security Assessment Report Automation Framework for Defense Closed Networks," Convergence Security Journal, Vol. 25, No. 4, pp. 31-43, Oct. 2025. DOI: 10.33778/kcsa.2025.25.4.031

[14] ISO/IEC 20000-1:2018, "Information Technology — Service Management — Part 1: Service Management System Requirements," ISO, Sep. 2018.

Authors



You-Min Kim received the B.S. degree in Information and Communication Engineering from Chosun University, Korea, in 2023. He is currently pursuing the M.S. degree in Information Technology Convergence at the Graduate School of Technology & Convergence, Kongju National University, Korea. He currently works in IT infrastructure and AI transformation in industry. He is interested in LLM agents, IT service management, retrieval-augmented generation, and AI security.



Jae-Woong Kim received the B.S. degree in Electronic Computer Science from Chung-Ang University, Korea, in 1983, the M.S. degree in Computer Engineering from Chung-Ang University, Korea, in 1988, and the Ph.D. degree in Computer Engineering from Daejeon University, Korea, in 2000. Dr. Kim joined the faculty of the Department of Software at Kongju National University, Korea, in February 1992. He is currently a Professor in the Department of Software at Kongju National University. He is interested in software engineering, big data, and multimedia engineering.



Won-Young Jeon received the B.S. degree in Health Science from Dongnam Health University, Korea, in 2019, and the Ph.D. degree from Kongju National University, Korea, in 2026. Dr. Jeon joined the faculty of the Department of Software at Kongju National University, Korea, in March 2026. She is currently a faculty member in the Department of Software at Kongju National University. She is interested in artificial intelligence, big data, LLM, and image processing.

www.kci.go.kr