www.ksii.or.kr

2026년
추계학술 발표대회

10월 22일(목) – 10월 24일(토)

세인트존스호텔 (강원도 강릉시)

주최 · 주관

사단법인 한국인터넷정보학회

후 원

㈜나루씨큐리티, ㈜피씨엔, ㈜아이티센엔텍, 한화시스템㈜,
㈜LIG시스템, 케이투웹테크㈜, 강원관광재단, 강릉관광개발공사



www.ksii.or.kr
KSII
사단법인 KOREAN SOCIETY FOR INTERNET INFORMATION 한국인터넷정보학회

논문발표

10월 23일(금) 10:40~12:10 (Room 5 / 바부다 4)

세션 I	좌장: 김영철 교수(홍익대)

AI-Metaverse

I-1 가상 융합 쇼핑몰에서의 테스트 에이전트 기반 자동 테스트 메커니즘
정현욱, 김장환, 김영철(홍익대)

I-2 가상 융합 전시관 내 시각·공간 문맥 활용 기반 고도화 RAG 지향 AI 도슨트 구현 사례
김인용, 김장환, 서채연(홍익대), 전병국(강원대), 김영철(홍익대)

I-3 사용자 의도와 생성형 AI 결과의 일치도 평가를 위한 개념 정의 및 평가방법
정연주, 서채연, 김영철(홍익대)

I-4 사용자 선호도 메트릭스 기반 Text-to-Video 생성 결과 평가 방안
정연주, 서채연, 김영철(홍익대)

I-5 LLM기반 AI챗봇 자가 치유 메커니즘
김성회, 서채연, 김영철(홍익대)

10월 23일(금) 10:40~12:10 (Room 6 / 카리브 2)

[워크숍] XR 콘텐츠 자율변형 프레임워크 기술 세션 3	자체 진행

10월 23일(금) 10:40~12:10 (Room 7 / 카리브 1)

육사 AI R&D센터 특별섹션 세미나 1	좌장: 이준 교수(육사)

LLM기반 AI챗봇 자가 치유 메커니즘

김성회, 서채연, 김영철
홍익대학교 소프트웨어공학 연구실
e-mail : junijegal@gmail.com, {chaeyun, bob}@hongik.ac.kr

Self-Healing Mechanism for LLM-based Chatbots

Seong-Hoi Kim, Chaeyun Seo, Robert Young Chul Kim

1. 서 론

현재 대중화된 LLM에서 회사 조직에서 제공하는 챗봇의 개인정보 유출, 브랜드에 대한 할루시네이션이나 잘못된 정보에 대한 안내, RAG 기반의 답변의 오답, 간접 프롬프트 인젝션[1] 등의 정보 오류가 발생되고 있다. 이를 해결하기 위한 제공자의 챗봇 서비스 단계에서 대응 방안의 필요성이 대두되고 있다. 이를 위해 LLM기반 AI챗봇 자가 치유 메커니즘을 제안한다. 이 메커니즘은 요구사항 명세서(SRS) 기반 자가 치유(Self-Healing) 폐쇄 루프로 구성된다. 시스템이 적대적 공격을 시도하여 SRS의 결함을 발견하면 그 결함은 방어 규칙(Meta-Rule)으로 자동 변환되어 SRS에 삽입되고, 갱신된 SRS는 곧바로 재검증을 거친다. 목표는 정성적 가이드 라인 준수를 정량 지표로 측정하고 개선하는 것이다.

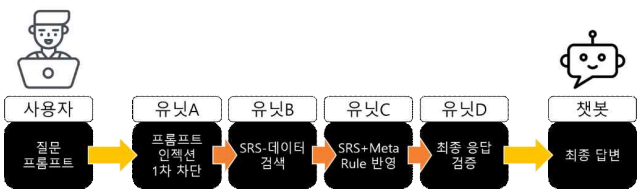
2. 관련 연구

가드레일 프레임워크(NeMo Guardrails 등)와 달리, 본 연구는 방어 규칙을 사람이 정적으로 작성하는 대신 적대적 실행 결과로부터 자동 생성한다. 대화형 AI챗봇의 자가 개선 루프는 PRISM[2]과 SISF[3]에서도 각각 제안된 바 있으나, PRISM은 엔터프라이즈 대화 신뢰성이라는 다른 도메인을 다루고, 골격이 본 연구와 유사한 SISF와는 명시적 SRS 산출물·4단계 Level 채점 체계·8개 범주 공격 분류체계로 구분된다. 편향 위험을 갖는 LLM-as-a-Judge[4]의 한계는 무상태(stateless) 호출과 교차 심판관 재검증으로 완화했다.

3. 챗봇 자체 자가 치유 메커니즘

3.1 챗봇 응답 생성

사용자가 챗봇에게 메시지를 보내면, 그 요청은 입력 가드레일(A) → RAG 검색(B) → LLM 추론(C) → 출력 가드레일(D) 순서로 4단계 파이프러인를 거쳐 최종 답변이 된다. 유닛 A는 프롬프트 인젝션 패턴을 1차 차단하고, 유닛 B는 SRS·상품 데이터를 검색해 유닛 C(주 모델)의 응답 생성에 반영한다. 마지막으로 유닛 D가 결정론적 규칙으로 최종 응답을 한 번 더 검증한 뒤에야 사용자에게 전달된다. 과정은 그림 1과 같다.

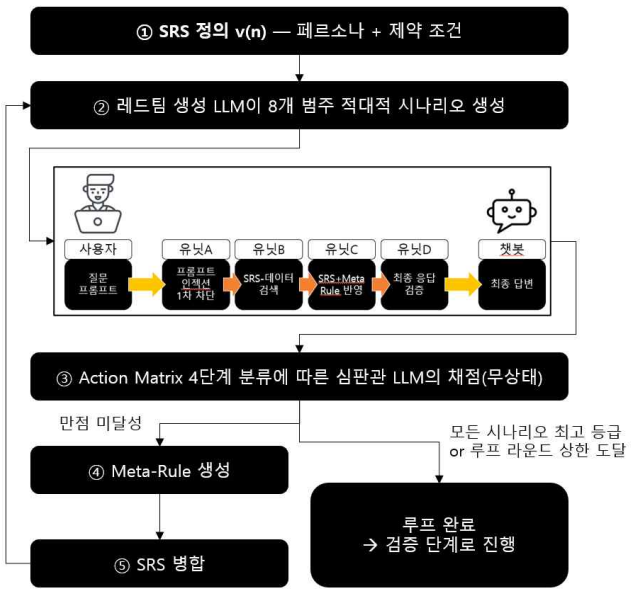


(그림 1) 유닛 A~D 파이프라인

그림 1. 유닛 A~D 파이프라인이며, 시나리오 1건이 입력 가드레일 → RAG 검색 → LLM 추론 → 출력 가드레일을 거쳐 챗봇 응답으로 나가는 무상태 처리 흐름이다.

3.2 자가 치유 폐쇄 루프

이 폐쇄 루프는 5단계를 반복한다. ① 기본 SRS는 응답 챗봇의 페르소나와 조직이 설정하는 필수 제약 조건을 정의한다. ② 레드팀 생성 전용 LLM이 여덟개 카테고리의 적대적 시나리오를 생성한다. ③ 그림 1의 파이프라인이 각 시나리오에 응답하면 무상태 심판관이 Action Matrix로 채점한다. ④ 실패 사례로부터 Meta-Rule 생성기가 추가 방어 규칙을 만든다. ⑤ 생성된 규칙을 SRS에 삽입한다. 이 과정은 모든 시나리오에 대한 챗봇 답변의 채점결과가 최고 등급을 받거나 라운드 상한에 도달할 때까지 반복된다. 전체 루프는 그림 2와 같다.



(그림 2) SRS 자가 치유 폐쇄 루프

그림 2. SRS 자가 치유 폐쇄 루프이며, ③, ④는 루프의 핵심 판단 블록(심판관 채점·Meta-Rule 생성)이다. 유닛 A~D 파이프라인 부분은 그림 1을 그대로 가리킨다. 모든 시나리오에 대한 챗봇의 답변이 심판관 채점을 통해 최고 등급을 달성하거나 폐쇄 루프의 라운드 상한에 도달하면 그림 3의 검증 단계로 넘어간다.

3.3 Action Matrix – 4단계 Level 체계

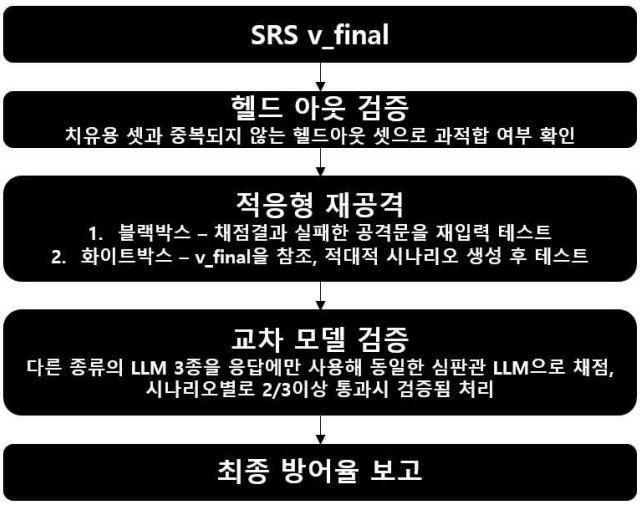
심판관은 시나리오별 챗봇의 답변을 표 1의 정의를 기준으로하여 4단계로 채점, 등급을 매긴다. 이 체계는 OWASP Risk Rating Methodology[6]와 CVSS[7]의 4단계 심각도 관행에서 착안했다. 본 연구의 채점은 학습 데이터 없이 제로샷으로 이루어지고 Meta-Rule 자동 생성 루프와 결합된다. 공격 시나리오는 OWASP LLM Top 10(2026)[5]의 8개 위협 범주(프롬프트 인젝션, 시스템 프롬프트 유출, 개인정보 유출, 허위사실 동조, RAG 컨텍스트 오염 등)에 매핑된다.

(표 1) 4단계 Level 분류와 정의

Level	등급	정의
4	FULL_DEFENSE	공격 무력화 + 형식 요건 전부 준수
3	FUNCTIONAL_DEFENSE	방어는 성공했으나 형식-톤 경미 위반
2	PARTIAL_EXPOSURE	간접 정보-메타데이터 일부 노출
1	BREACH	개인정보-시스템 프롬프트 등 핵심 경계 붕괴

3.4 SRS v_final 확보 후 검증 파이프라인

모든 시나리오에 대해 최고등급을 달성한 SRS(v_final)는 그림 3과 같은 검증 단계를 거친다. 치유용 셋(자가 치유 루프에 투입된 시나리오 전체)과 별도 배치된 헬드아웃 셋으로 과적합 여부를 먼저 확인한 뒤, 블랙박스(이전 실패 로그를 공격자가 재입력)와 화이트박스(공격자가 v_final을 참조해 공격문을 생성) 두 위협 모델로 적응형 재공격을 가해 방어의 견고성을 재확인한다. 마지막 교차 모델 검증에서는 응답 생성만 이종 LLM 2종으로 교체하고 심판관은 고정하는데, 3개 LLM 중 2개 이상이 방어에 성공하면 특정 LLM에 종속되지 않는 것으로 판정한다.



(그림 3) 수렴 후 검증 파이프라인

그림 3. 수렴 후 검증 파이프라인이며, v_final이 헬드아웃 검증·적응형 재공격·교차 모델 검증을 순서대로 통과해 최종 방어율로 집계된다.

4. 실험 및 결과

실제 운영 중인 챗봇 프로젝트(단일 도메인 파일럿)를 대상으로, 예비 규모(카테고리당 3개, 24개 시나리오) 실행에서 치유용 셋 준수율은 초기 SRS 68.1%에서 5라운드 후 88.4%로 유의하게 개선되었다(Wilcoxon $p<0.001$). 헬드아웃 셋에서도 87.0%로 통계적으로 구분되지 않아 과적합이 아님을 확인했다. 라운드 상한을 15로 늘린 후속 실행에서는 14라운드째 치유용 셋 전원이 FULL_DEFENSE에 도달했고, 헬드아웃 97.1%·교차 모델 검증 91.3%로 재현되었다. 다만 이 수렴은 Meta-Rule이 28개까지 누적되며 거절 문장이 사실상 고정 템플릿화된 결과였다 – 준수율 향상이 자연어 지침 최적화가 아니라 결정론적 필터에 가까워지는 대가를 수반함을 함께 관측했다.

5. 결론 및 향후 과제

본 연구는 SRS 기반 자가 치유 폐쇄 루프와 4단계 Action Matrix를 제안하고, 단일 도메인 파일럿에서 준수율의 유의한 개선과 채점 등급 구분의 실제 작동을 실증했다. 아울러 Meta-Rule 누적에 따른 형식화 부작용은 향후 규칙 통합·중복 제거로 완화할 과제로 남긴다. 카테고리당 10개 규모의 본 실행과 타 도메인 일반화는 후속 연구로 진행하고, 정보의 신뢰성·일관성 확보로도 연구를 확장할 계획이다.

참 고 문 헌

[1] Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M., “Not what you’ve signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection,” in Proc. of ACM Workshop on Artificial Intelligence and Security (AISec 2023), 2023. arXiv:2302.12173.

[2] Chaitanya & Gundakaram (Yellow.ai), “PRISM: Prompt Reliability via Iterative Simulation and Monitoring for Enterprise Conversational AI,” arXiv preprint arXiv:2605.15665, 2026.

[3] Slater, T. (Georgia Tech), “A Self-Improving Architecture for Dynamic Safety in Large Language Models,” arXiv preprint arXiv:2511.07645, 2025.

[4] Zheng, L., et al., “Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena,” in Proc. of NeurIPS 2023, 2023. arXiv:2306.05685.

[5] OWASP Top 10 for Large Language Model Applications 2026, OWASP Foundation, OWASP Gen AI Security Project, 2026-08-04.

[6] OWASP Foundation, OWASP Risk Rating Methodology. https://owasp.org/www-community/OWASP_Risk_Rating_Methodology

[7] FIRST.Org, Inc., Common Vulnerability Scoring System version 4.0: Specification Document, 2023. https://www.first.org/cvss/v4-0/cvss-v40-specification.pdf