생성형 AI 문서 범람에 따른 지식오염 대응 프레임워크 연구

조소연*, 정지호**, 이주찬**, 이용준***
*극동대학교 인공지능보안학과
**극동대학교 해킹보안학과
***극동대학교 해킹보안학과 교수
e-mail: seeujfg834383@gmail.com

A Framework for Mitigating Knowledge Pollution Caused by the Proliferation of Generative AI Documents

SoYeon Jo*, JiHo Jeong**, JooChan Lee**, Yongjun Lee***
*Dept. of AI Security, Far East University
**Dept. of Hacking Security, Far East University
***Professor, Dept. of Hacking Security, Far East University

요 약

본 논문에서는 생성형 AI 기술 확산에 따라 보고서, 회의록, 교육자료, 매뉴얼 등 다양한 문서가 대량으로 생성·유통되는 환경에서 발생할 수 있는 지식오염(Knowledge Pollution) 문제를 분석했다. 생성형 AI가 생성한 문서는 높은 생산성과 효율성을 제공하지만, 사실 오류 정보, 중복·유사 문서, 출처 불명 콘텐츠, 최신성 저하 정보 등이 지속적으로 축적될 경우 지식체계의 신뢰성과 활용성을 저하시킬 수 있다. 이에 따라 생성형 AI 시대의 새로운 정보 품질 관리 필요성이 제기된다. 본 연구는 생성형 AI 문서의 생애주기를 생성-식별-검증-저장-운영의 5단계로 분석하고, ISO 정보 품질 관리 원칙의 예방-검증-통제-개선 원칙을 설계 기반으로 삼아 지식오염 대응 프레임워크를 제안했다. 제안 프레임워크는 입력 통제 계층, AI 문서 식별 계층, 품질 검증 계층, 저장·검색 관리 계층, 모니터링 및 개선 계층의 5단계 구조로 구성되며, 생성형 AI 문서의 신뢰성과 활용 가치를 높이는 것을 목표로 한다. 또한 제안 프레임워크는 기업, 교육기관, 연구기관, 온라인 플랫폼 등 다양한 환경에서 적용 가능하도록 설계되었다. 본 연구는 생성형 AI 문서를 단순 저장 대상이 아닌 관리 대상 정보자산으로 재정의하고, 생성형 AI 시대의 지식관리 거버넌스 체계 수립에 기여한다.

주제어: 생성형 AI, 지식오염, 정보품질관리, 정보 신뢰성, AI 거버넌스

1. 서론

생성형 인공지능(Generative AI)의 확산으로 보고서, 회의록, 매뉴얼, 질의응답 자료 등 다양한 문서가 자동 생성되고 있으며, 기업·교육기관·공공기관 등 여러 조직에서 이를 업무 효율화 수단으로 적극적으로 활용하고 있다. McKinsey의 2025년 조사에 따르면, 다수의 조직이 이미 하나 이상의 업무 기능에서 AI를 활용하고 있는 것으로 나타났으며, 생성형 AI는 문서 생산성과 정보 접근성을 높이는 핵심 기술로 평가되고 있다[1].

생성형 AI가 생성한 문서는 사실 오류, 출처 불명 정보, 중복 콘텐츠, 최신성 저하 등의 한계를 가질 수 있다. 이러한 문서가 조직의 내부 문서 저장소 등에 지속적으로 축적될 경우 정보의 정확성과 신뢰성이 저하될 수 있다. NIST는 생성형 AI 활용 시 콘텐츠 출처 검증, 위험 관리, 지속적 모니터링 체계의 필요성을 강조하고[2], OECD 또한 생성형 AI가 허위정보 생성 및 정보 왜곡 위험을 동반할 수 있다고 지적했다[3].

생성형 AI 기반 문서가 대량 생산·저장·재활용되는 환경에서 저품질 정보가 지식체계 내부에 누적되는 '지식오염(Knowledge Pollution)' 문제가 발생할 수 있다. 지식오염은 사실 오류 정보의 축적, 중복·유사 문서 증가, 출처 불명 정보 확산, 최신성 저하 정보의 장기 잔존 등을 포함하며, 이는 검색 정확도 저하, 의사결정 오류, 업무 효율성 감소 등 다양한 부정적 결과로 이어질 수 있다[4].

본 연구는 생성형 AI가 만든 문서 범람으로 인해 발생하는 지식오염의 주요 유형과 위험요인을 분석하고, 완화하기 위한 대응 프레임워크를 제안하고자 한다.

2. 생성형 AI 시대 문서 범람과 지식오염 문제 분석

대규모 언어 모델 기반 서비스는 사용자의 지시문만으로 짧은 시간 안에 완성도 높은 텍스트를 생성할 수 있어 문서 생산 비용 절감과 업무 효율성 향상 수단으로 주목받고 있다

[5].

그러나 생성형 AI 문서의 급격한 증가는 정보 품질 저하라는 새로운 위험을 동반한다.

첫째, 사실 오류 정보의 축적 문제이다. 생성형 AI는 문맥상 자연스럽지만 실제로는 부정확한 내용을 포함한 문서를 생성할 수 있다. Stanford University 인간중심 AI 연구소는 최신 보고서에서 생성형 AI의 성능 향상에도 불구하고 사실성(factuality)과 신뢰성 문제는 여전히 핵심 과제로 남아 있다고 평가했다[6].

둘째, 중복·유사 문서 증가 문제이다. 생성형 AI는 동일한 주제에 대해 표현만 일부 변경된 문서를 손쉽게 대량 생산할 수 있다. 이로 인해 조직 내 문서 저장소에는 내용상 유사한 자료가 빠르게 증가할 수 있으며, 검색 결과의 정확성과 정보 탐색 효율성이 저하될 수 있다.

셋째, 출처 불명 정보 확산 문제이다. 생성형 AI 문서는 작성 근거, 원자료, 참고 출처가 명확히 제시되지 않는 경우가 많다. 유럽연합 집행위원회(European Commission)는 AI 생성 콘텐츠 활용 시 투명성과 추적 가능성의 중요성을 강조하였으며, 생성물에 대한 출처 표시와 책임성 확보가 필요하다고 제시했다[7]. 출처가 불명확한 문서가 반복적으로 활용될 경우 정보 검증은 더욱 어려워질 수 있다.

넷째, 최신성 저하 정보의 잔존 문제이다. 생성형 AI는 특정 시점까지의 데이터를 기반으로 학습되므로 이후 변경된 법률, 정책, 기술 환경이 즉시 반영되지 않을 수 있다.

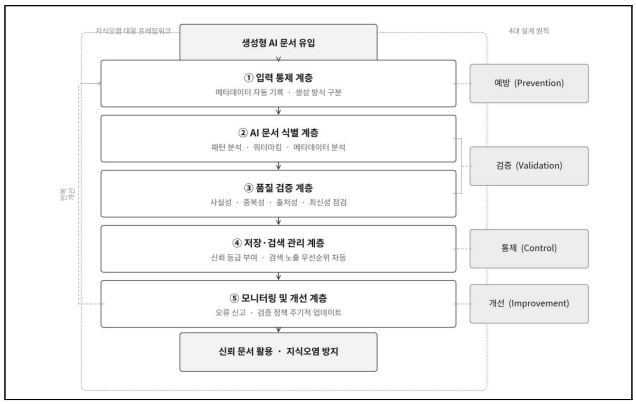
이와 같은 문제들이 지속적으로 누적될 경우 단순한 정보 과잉을 넘어 저품질 정보가 지식체계 내부에 축적되는 '지식오염' 현상이 발생한다. 지식오염은 정보의 정확성, 검색 가능성, 활용 가능성, 신뢰성을 동시에 저하시킬 수 있으며, 장기적으로는 조직 경쟁력 저하와 사회적 정보 신뢰도 약화로 이어질 수 있다.

3. 생성형 AI 문서 기반 지식오염 대응 프레임워크 제안

생성형 AI가 생산한 문서의 대량 유입은 사실 오류 정보 축적, 중복 문서 증가 등 다양한 형태의 지식오염 문제를 유발할 수 있다. 따라서 단순히 생성형 AI 사용을 제한하는 방식보다는, 생성된 문서의 생애주기 전반을 관리할 수 있는 체계적 대응 구조가 필요하다. 이에 본 연구는 생성형 AI 문서의 생성, 저장, 검색, 재사용 전 과정을 통합 관리하기 위한 지식오염 대응 프레임워크를 제안한다.

제안 프레임워크는 문헌 분석과 기존 정보 품질 관리 모델을 기반으로 설계되었다. 구체적으로, 본 연구는 생성형 AI 문서의 생애주기(document lifecycle)를 생성–식별–검증–저장–운영의 5단계로 구분하고, 각 단계에서 지식오염이 발생할 수 있는 위험 지점을 식별했다. 이를 토대로 각 위험 지점에 대응하는 관리 계층을 순차적으로 설계하였으며, 계층 간 의존성과 처리 순서를 고려하여 선형 파이프라인(linear pipeline) 구조를 채택했다.

프레임워크의 4대 설계 원칙은 국제표준화기구(ISO)의 정보 품질 관리 표준(ISO 8000) 및 AI 관리체계 표준(ISO/IEC 42001)에서 제시하는 핵심 관리 원칙을 참조하여 도출했다[8][9]. 예방(Prevention) 원칙은 문서 유입 단계의 메타데이터 통제, 검증(Validation) 원칙은 AI 문서 식별 및 품질 점검 단계, 통제(Control) 원칙은 저장·검색 결과의 신뢰도 관리, 개선(Improvement) 원칙은 운영 피드백을 통한 정책 고도화에 각각 대응한다. 또한 세계 주요 기관들이 강조하는 AI 투명성, 책임성, 지속 모니터링 요구사항을 반영하여 실무 적용 가능성을 높이고자 했다[10].



[그림 1] 생성형 AI 문서 기반 지식오염 대응 프레임워크 구조도

3.1 입력 통제 계층

입력 통제 계층은 생성형 AI가 만든 문서가 조직 내부 시스템으로 유입되는 초기 단계에서 작동한다. 해당 계층에서는 문서 등록 인터페이스가 사용자의 저장 요청을 수신하면, 메타데이터 수집 모듈을 호출하여 생성 방식(인간 작성/AI 작성/혼합 작성), 사용 모델명, 작성 일시, 작성 목적을 자동 추출하고 문서 헤더에 태깅한 후 다음 계층으로 전달한다. Microsoft는 기업용 AI 도입 환경에서 생성물의 추적 가능성과 관리 로그 확보가 중요하다고 제시하였으며, 이는 사후 검증 및 책임성 확보의 기초 자료가 된다[11].

3.2. AI 문서 식별 계층

두 번째 계층은 문서가 생성형 AI에 의해 작성되었는지 또는 일부 수정되었는지를 식별하는 단계이다. 최근 AI 생성 텍스트 탐지 기술은 완전한 정확도를 보장하지는 않지만, 일정 수준의 식별이 가능하다. 이에 따라 식별 모듈은 전 계층에서 전달받은 문서를 입력으로 받아 텍스트 패턴 분석기·워터마크 검출기·메타데이터 파서를 병렬로 구동하고, 각 결과를 종

합하여 AI 생성 확률 점수(0~1)를 산출한다. 이 점수는 문서 속성으로 저장되어 다음 품질 검증 계층의 판단 기준으로 활용된다. Google DeepMind는 AI 생성 콘텐츠 식별 기술이 향후 정보 생태계 관리에 중요한 역할을 할 것으로 전망했다[12].

3.3. 품질 검증 계층

품질 검증 계층은 지식오염 방지를 위한 핵심 단계로서, 생성 문서의 사실성, 중복성, 출처성, 최신성을 점검한다. 품질 검증 모듈은 사실성 검증기·중복성 검사기·출처 파서·최신성 검사기를 순차 실행한다. 사실성 검증기는 외부 신뢰 데이터베이스 및 내부 기준 문서와 비교하여 오류 여부를 판단하며, 중복성 검사기는 기존 저장 문서와의 코사인 유사도를 산출하여 중복 여부를 식별한다. 출처 파서는 참조 URL 및 출처 표기 유무를 확인하고, 최신성 검사기는 문서 내 날짜와 최신 정책 기준일을 대조한다. 각 검사 결과는 통과/경고/반려로 분류되며, 반려 판정 시 해당 문서는 작성자에게 피드백과 함께 반환된다. Deloitte는 생성형 AI 도입 시 Human-in-the-loop 검증 절차를 병행해야 조직 리스크를 줄일 수 있다고 제시했다[13].

3.4. 저장·검색 관리 계층

검증을 통과한 문서는 저장·검색 관리 계층으로 이동한다. 이 단계에서 등급 산정 모듈은 품질 검증 계층의 결과를 입력받아 사실성·중복성·출처성·최신성 점수를 가중 합산하여 신뢰 등급(High/Medium/Low)을 결정하고 문서 메타데이터에 기록한다. 검색 엔진은 사용자 질의 시 신뢰 등급을 랭킹 가중치로 반영하며, Low 등급 문서에는 '검증 미완료' 레이블을 노출한다. 각 점수의 가중치는 조직의 정책 목표에 따라 조정 가능하도록 설계되었다. Amazon Web Services는 생성형 AI를 활용한 지식관리 환경에서 자연어 기반 검색과 문서 자동 업데이트를 통해 정보 접근성과 검색 정확도를 높일 수 있다고 제시했다[14].

3.5. 모니터링 및 개선 계층

마지막 계층은 프레임워크 운영 이후 지속적 개선을 담당한다. 모니터링 모듈은 사용자 오류 신고, 문서 활용률, 검색 만족도, 잘못된 인용 빈도를 수집하여 품질 지표 대시보드에 집계한다. 정책 관리자는 이 대시보드를 기반으로 품질 검증 계층의 판단 임계값과 등급 산정 가중치를 주기적으로 조정하며, 변경된 정책은 각 계층에 자동 반영된다. PwC는 생성형 AI 거버넌스의 핵심 요소로 지속적 성과 측정과 리스크 모니터링을 제시했다[15].

4. 적용 시나리오 및 기대효과

제안한 생성형 AI 문서 기반 지식오염 대응 프레임워크가 실제 환경에서 어떻게 활용될 수 있는지 주요 적용 시나리오를 제시하고, 이를 통해 기대되는 효과를 분석하고자 한다.

4.1. 조직 지식관리시스템(KMS) 적용 시나리오

첫 번째 적용 시나리오는 기업 및 기관의 지식관리시스템 환경이다. 최근 많은 조직은 업무 매뉴얼, 프로젝트 보고서, 내부 규정, 고객 응대 사례 등을 중앙 저장소에 축적하고 있으며, 여기에 생성형 AI를 활용한 문서 자동 생성 기능이 결합되고 있다. 그러나 검증되지 않은 AI 생성 문서가 대량 등록될 경우 검색 정확도 저하와 업무 혼선을 초래할 수 있다.

4.2. 교육·연구기관 문서관리 적용 시나리오

두 번째 적용 시나리오는 대학, 연구소, 교육기관 등 지식 생산 중심 조직이다. 생성형 AI는 강의자료 초안, 연구보고서 정리, 행정문서 작성 등에 활용될 수 있으나, 검증되지 않은 자료가 학습 또는 연구 자료로 활용될 경우 학문적 신뢰성과 연구 무결성에 영향을 줄 수 있다.

4.3. 온라인 정보 플랫폼 적용 시나리오

세 번째 적용 시나리오는 온라인 커뮤니티, 콘텐츠 플랫폼, 정보 포털 등 공개형 정보 유통 환경이다. 최근 생성형 AI를 활용한 블로그 글, 상품 설명, 뉴스 요약, 후기 콘텐츠 등이 급증하고 있으며, 일부는 사실 검증 없이 배포되고 있다. 이러한 콘텐츠가 누적될 경우 사용자들은 신뢰 가능한 정보를 식별하기 어려워진다.

4.4. 기대효과

제안 프레임워크의 적용을 통해 다음과 같은 효과를 기대할 수 있다.

첫째, 정보 신뢰도 향상이다. AI 생성 문서에 대한 검증과 등급 분류를 통해 사용자는 신뢰성 높은 정보를 우선 활용할 수 있다.

둘째, 검색 정확도 개선이다. 중복 문서 제거와 메타데이터 기반 관리가 가능할 경우 정보 탐색 시간이 단축되고 검색 만족도가 향상될 수 있다.

셋째, 의사결정 오류 예방이다. 사실 오류 또는 최신성 저하 문서의 활용 가능성을 낮춤으로써 조직 내 전략적·운영적 판단의 정확성을 높일 수 있다.

넷째, 지속 가능한 AI 활용 기반 조성이다. 단순히 생성형 AI 사용을 제한하는 것이 아니라, 관리가 가능한 활용 구조를 마련함으로써 생산성과 신뢰성을 동시에 확보할 수 있다.

5. 결론

생성형 AI 기술의 확산은 다양한 조직과 플랫폼에서 문서 생산성을 높이는 동시에, 검증되지 않은 정보가 대규모로 유통·축적되는 새로운 구조적 위험을 동반하고 있다. 본 연구는 이러한 위험을 단순한 정보 과잉이나 콘텐츠 품질 문제로 보지 않고, 저품질 정보가 지식체계 내부에 누적되어 검색·의사결정·학습 전반에 영향을 미치는 '지식오염(Knowledge Pollution)' 현상으로 정의했다. 이는 기존 문서관리나 정보보안 연구에서 본격적으로 다루어지지 않은 새로운 문제 영역으로, 생성형 AI 시대에 요구되는 지식관리 패러다임의 전환을 촉구한다는 점에서 본 연구의 첫 번째 기여가 있다.

이를 해결하기 위해 본 연구는 생성형 AI 문서의 생성부터 저장, 검색, 활용, 사후 관리까지 전 생애주기를 포괄하는 5계층 대응 프레임워크를 제안했다. 제안 프레임워크는 입력 통제, AI 문서 식별, 품질 검증, 저장·검색 관리, 모니터링 및 개선의 각 계층이 ISO 정보 품질 관리 원칙의 예방–검증–통제–개선 원칙에 체계적으로 대응하도록 설계되었다. 기존 문서관리 시스템이 문서를 수동적 저장 대상으로 취급하는 데 반해, 본 프레임워크는 생성형 AI 문서를 능동적으로 관리해야 할 정보자산으로 재정의한다는 점에서 기존 접근과의 차별성을 가진다.

또한 제안 프레임워크는 특정 기술 솔루션에 의존하지 않는 절차적·정책적 구조로 설계되어, 기업 지식관리시스템, 교육·연구기관, 온라인 정보 플랫폼 등 다양한 환경에 범용적으로 적용 가능하다. 이는 단일 조직이나 기술 환경에 한정된 해결책이 아닌, 생성형 AI 도입 환경 전반에 적용할 수 있는 거버넌스 모델을 제시한다는 점에서 실무적 기여를 갖는다.

다만 본 연구는 개념적 프레임워크 제안에 집중한 연구로서, 실제 조직 환경에서의 정량적 성능 검증과 산업별 세부 적용 방안 도출에는 한계가 있다. 향후 연구에서는 프레임워크를 특정 조직에 시범 적용하여 정보 신뢰도 향상 수준, 검색 효율성 개선 정도, 운영 비용 대비 효과를 실증적으로 분석할 필요가 있다. 나아가 생성형 AI 기술의 발전에 따라 새롭게 등장하는 지식오염 유형을 지속적으로 식별하고 프레임워크를 고도화하는 연구도 요구된다.

참고문헌

[1] McKinsey & Company, “The State of AI: Global Survey 2025,” McKinsey Global Institute, 2025.

[2] National Institute of Standards and Technology (NIST), “Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile,” NIST AI 600-1, 2024.

[3] Lorenz, P., K. Perset and J. Berryhill (2023), "Initial policy considerations for generative artificial intelligence," OECD Artificial Intelligence Papers, No. 1, OECD Publishing, Paris. doi:10.1787/fae2d1e6-en.

[4] B. Mitra et al., “Sociotechnical Implications of Generative Artificial Intelligence for Information Access,” arXiv preprint arXiv:2405.11612, 2024.

[5] Accenture, "Reinventing Enterprise Operations with Gen AI," Accenture Research, 2024.

[6] Stanford Institute for Human-Centered Artificial Intelligence, “AI Index Report 2025,” Stanford University, 2025.

[7] European Commission, "Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (AI Act)," Official Journal of the European Union, 2024.

[8] International Organization for Standardization (ISO), “ISO 8000 Data Quality Standards,” 2024.

[9] International Organization for Standardization (ISO), “ISO/IEC 42001 Artificial Intelligence Management System,” 2024.

[10] UNESCO, “Recommendation on the Ethics of Artificial Intelligence,” 2021.

[11] Microsoft, "Responsible AI Transparency Report," Microsoft Corporation, May 2024.

[12] Google DeepMind, "Watermarking AI-generated text and video with SynthID," Google DeepMind Blog, May 2024.

[13] Deloitte, "State of Generative AI in the Enterprise," Deloitte AI Institute, 2024.

[14] Amazon Web Services (AWS), "Generative AI Use Cases for Knowledge Management," AWS Prescriptive Guidance, 2024.

[15] PwC, "2024 US Responsible AI Survey," PwC United States, 2024.