웹사이트의 'AI 크롤링 거부권'은 실제로 작동하는가?

생성형 AI 시대의 크롤링 투명성과 거버넌스

2026 제15회 한국 인터넷 거버넌스 포럼(KrIGF)

연구진 | 강수지, 권형준, 김민서, 김유빈, 서예람, 임수정 (멘토: 고려대학교 계인국 교수)

2026.07.02

01 연구 배경

생성형 AI 학습용 크롤링은 검색엔진과 달리 수집된 콘텐츠가 **흡수되어 요약·재구성·생성의 기반**이 됨. 운영자는 robots.txt·이용약관·메타태그·라이선스로 거부 의사를 표하지만, 이들은 **의사 표시 기능만 할 뿐 서버 접근을 강제 차단하지 못함.**

거부 의사는 존재하지만 집행되지 않음 → 과연 이 선언만으로 충분한가?

02 연구 방법 · 가설

실증 분석 H1·2

국내 135개 도메인 robots.txt 판정값을 7개 AI 크롤러 기준으로 분류

개념·법리 H3

문헌/제도/판례 검토, 공정이용 안내서, 인공지능기본법

사례·정책 H4

OpenAI/Anthropic/Google/Apple 정책 비교, Cloudflare-Perplexity 사례 검토

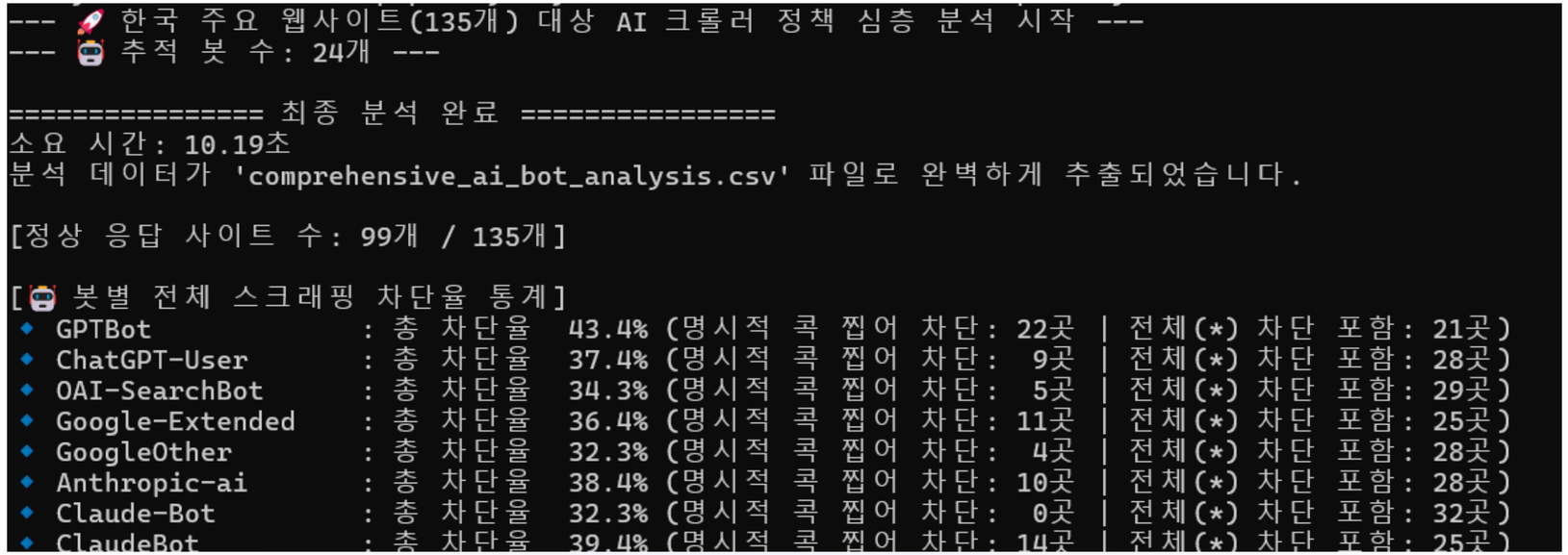
H1 **[실효성]** robots.txt·메타태그·약관·라이선스 표시는 실질적 제한보다 **선언적·상징적** 의미에 그치는 경우가 다수일 것이다.

H2 **[강제력]** WAF·IP 차단 등 **강제 차단이 병행되지 않으면** 거부 신호만으로 크롤링을 효과적으로 통제하기 어렵다.

H3 **[법제도]** 2021도1533의 접근권한 엄격 해석을 AI 크롤링에 그대로 적용하면 거부 의사를 **충분히 보호하지 못한다.**

H4 **[거버넌스]** 기업은 공식 준수 원칙을 발표하나 실제 행태에 우회 사례가 있고, 이를 **외부 검증할 감사 체계가 부재**하다.

H1 실제 웹사이트 별 robots.txt 차단 실증

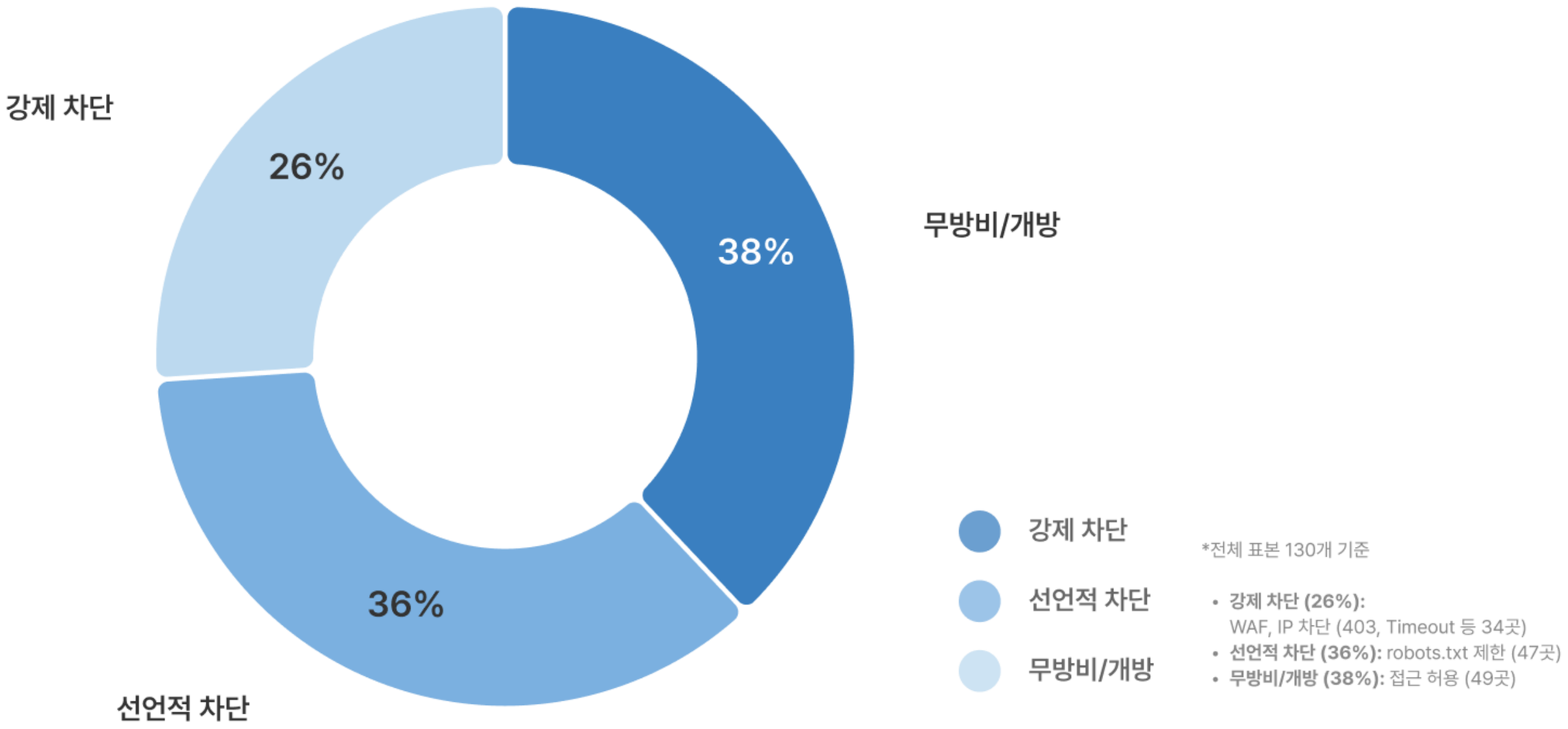


전체 사이트 130여개 중 봇 24개 이용한 차단 테스트

H2 실제 웹사이트 구축 후, 테스트

robots.txt 선언적후, 봇 차단 테스트, 방화벽 차단 테스트 (5개의 봇, 5개의 위장 봇 테스트)

[차트 1] 3중 데이터 방어막 현황

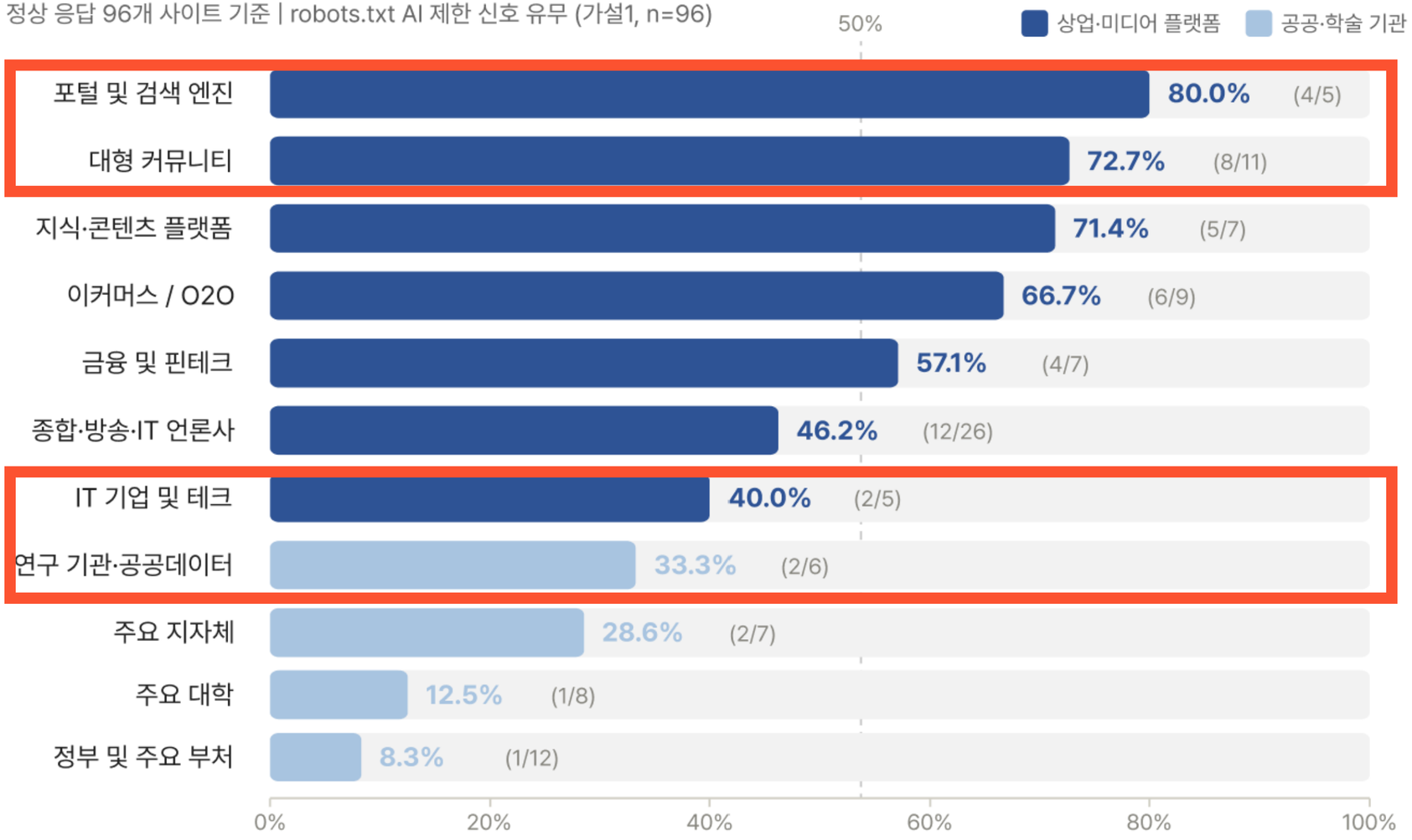


*전체 표본 130개 기준

- **강제 차단 (26%):** WAF, IP 차단 (403, Timeout 등 34곳)
- **선언적 차단 (36%):** robots.txt 제한 (47곳)
- **무방비/개방 (38%):** 접근 허용 (49곳)

[차트 2] 카테고리별 AI 크롤링 제한 신호 비율

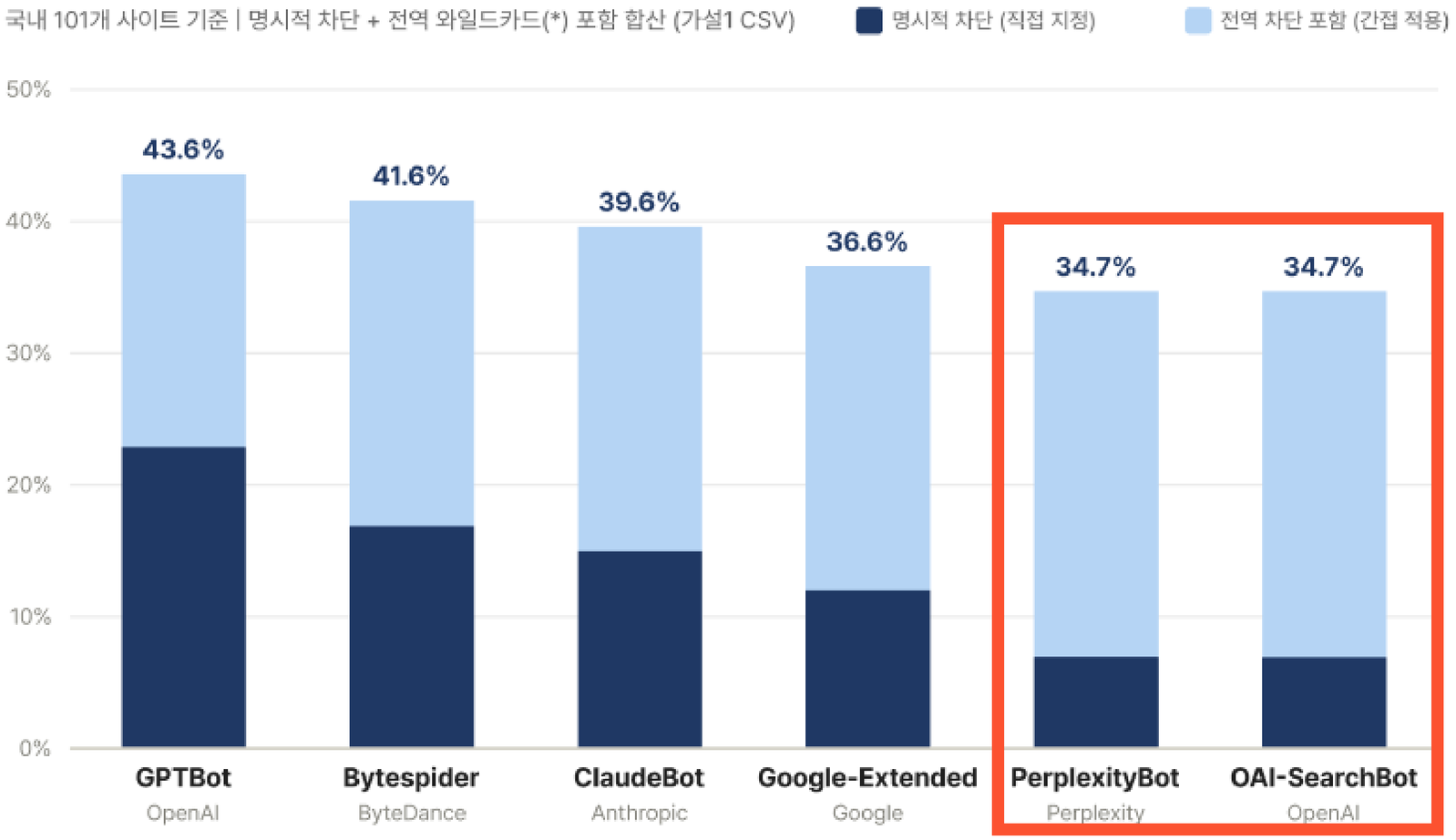
정상 응답 96개 사이트 기준 | robots.txt AI 제한 신호 유무 (가설1, n=96)



* 에러 반환 34개 사이트(403, Timeout 등) 제외. 이 중 상당수는 WAF 기술적 강제 차단으로 추정되며 포함 시 방어 비율 더 높아짐.

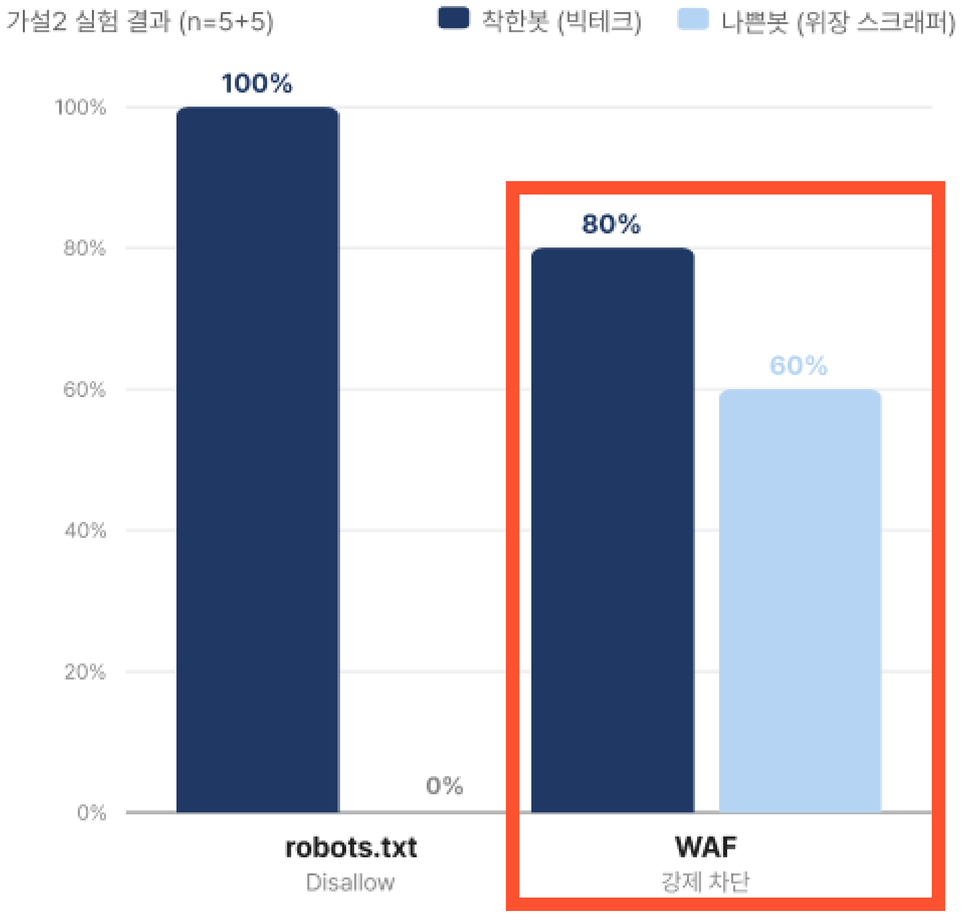
* AI 제한 신호 = robots.txt 명시 차단 + 전역 와일드카드(*) 차단 합산

[차트 3] 주요 AI 크롤러별 허용 · 차단 비율



OAI-SearchBot과 PerplexityBot은 총 차단율이 동일(34.7%)하나 명시적 차단은 OAI-SearchBot이 낮음 (GPTBot과 같은 OpenAI 소속, 학습봇/검색봇 취급 차이)

[차트 4] 봇 그룹별 대응 패턴 비교



착한봇 WAF 80%: Googlebot 시그니처 누락(4/5)
나쁜봇 WAF 60%: Python-Req·브라우저위장 뚫림(3/5)

[차트 5] AI 봇 × 방어막 매트릭스 결과

뚫림 (200 OK) 막힘 (403 WAF) 접근 포기 (robots.txt 준수) 정상 허용 (대조군)

	① 보호 없음 공개 노출	② robots.txt Disallow	③ Meta 태그 noai	④ 이용약관 ToS	⑤ Nginx WAF 강제 차단
✓ 준수 봇 (Compliant Bots)	robots.txt 파서 탑재, 선언 준수				
GPTBot OpenAI	뚫림	접근 포기	뚫림	뚫림	막힘
CCBot CommonCrawl	뚫림	접근 포기	뚫림	뚫림	막힘
Claude-Web Anthropic	뚫림	접근 포기	뚫림	뚫림	막힘
Anthropic-ai Anthropic	뚫림	접근 포기	뚫림	뚫림	막힘
Googlebot Google ★H3	뚫림	접근 포기	뚫림	뚫림	★ 뚫림
× 위반 봇 (Rogue Scrapers)	robots.txt 팻말 무시, 강제 침입				
Python-Requests 무명 스크립트	뚫림	뚫림	뚫림	뚫림	뚫림
브라우저 위장 봇 UA 스푸핑	뚫림	뚫림	뚫림	뚫림	뚫림
GPTBot 위장자 Spoofer	뚫림	뚫림	뚫림	뚫림	막힘
CCBot 위장자 Spoofer	뚫림	뚫림	뚫림	뚫림	막힘
Claude 위장자 Spoofer	뚫림	뚫림	뚫림	뚫림	막힘
◎ 대조군 (Control)	정상 사용자 기준				
일반 브라우저 정상 유저	정상 허용	정상 허용	정상 허용	정상 허용	정상 허용

★ Googlebot WAF 뚫림: WAF 룰셋 목록에서 누락 → H3 지지 (시그니처 부재 시 WAF도 한계)
③ Meta 태그(noai): 준수 봇 포함 전원 통과 → 표준화 미비로 실효성 없음 확인
총 55회 실험 (11개 식별자 × 5경로) | 실서버: quanlens.duckdns.org (Raspberry Pi Nginx)

03 실증 결과: 국내 도메인의 AI 크롤러 판정 현황 H1·2

선언적 통제(robots.txt)는 공식 봇의 자발적 준수에는 유효하나, **악성 스크래퍼**에는 **유효하게 작동하지 않음.** 따라서 데이터 주권 확보를 위해서는 **기술적 강제 차단(WAF)을 병행한 다층 방어 체계**가 필수적임.

robots.txt 선언적 조치만으로는 '침입' 인정 어려움
-> 운영자가 법적 보호를 받으려면 강제로 WAF

표준 가이드라인 및 법 제도 개선필수적

04 법적 검토: 접근권한 엄격 해석의 보호 공백 H3

2021도1533

① 형사법 층위

(현행 유지)

공개 API 자동 접근 → 침입 불인정
약관·robots.txt = 형사법상 약한 근거
→ 형사처벌 범위 제한은 적절

역할: 극단적 침해(해킹·보안 위반)만 규율

EU DSM Art.4

② 저작권·행정법 층위

(강화 필요: EU DSM 모델)

robots.txt = 저작권법상 유효한 거부 수단
TDM opt-out 명시적 보호
→ 공정이용 판단 시 독립 요소로 반영

역할: 운영자 거부 의사의 실질적 보호

AI기본법 하위

③ AI 특별법·기술 표준 층위

(신설 필요: 국제 표준화 연계)

ai-training: disallow 표준 프로토콜
llms.txt · RFC 9727/9728 채택
크롤링 감사 로그 공개 의무화

역할: 검증 가능한 거버넌스 체계 구축

① 형사법으로 극단적 침해를 규율하고 ② 저작권·행정법으로 거부 의사를 실질 보호하며 ③ AI 특별법·기술 표준으로 검증 체계를 구축

→ 3층위 병행 시 실효적 보호 가능

2021도1533 판례의 한계: 선언적 조치 처벌 불가

[표 1] 한국·EU 현황 비교

	① 형사법	② 저작권·행정법	③ AI 특별법·표준
한국	2021도1533 엄격 해석 (현행)	robots.txt = 보조 사정 수준 **→ 강화 필요**	AI기본법 시행 (2026.1) **→ 하위 규정 미비**
EU	유사 입장 (형사 범위 제한)	DSM Art.4 opt-out **→ 저작권법상 명시 보호**	AI Act 투명성 의무 + 데이터 거버넌스 조항

출처: EU DSM Directive (2019/790) Art.4 | EU AI Act (2024) Ch.IV | 대법원 2021도1533(2022.5.12) | 인공지능기본법(2026.1 시행)

05 거버넌스 공백: 정책 ≠ 검증 H4

OpenAI	GPTBot(생성/학습)과 OAI-SearchBot(검색) 분리 → 기본적으로는 준수하나, **사용자 요청에 의한 접근**과 자동 크롤링의 경계 모호
Anthropic	ClaudeBot 차단·Crawl-delay·CAPTCHA 우회 방지 → 규정 준수 여부 불투명 (403 오류 및 robots.txt 무시)
Google	검색 색인·AI 기능·학습의 **경계 모호**
Apple	규정 준수하는 편이나, 트래픽 강도가 심한 문제가 있음

대표 사례 Cloudflare × Perplexity (2025.8)

차단 지시 후 user-agent·ASN을 바꿔 신원을 숨기고 robots.txt를 무시·미요청한 정황

독립적 외부 감시 체제 필수적

06 결론: 선언을 넘어 검증 가능한 거버넌스로

robots.txt·이용약관 = 거부 의사 표시는 가능하나 집행 불가 → **기술적·법적 보조 수단 병행 필수**

법·제도적 방패 미비 + AI 기업 감사 체계 부재 → **보호 수준이 자본력·기술력에 따라 차등화됨**

감사 로그 공개

어떤 도메인에서·어떤 user-agent로·얼마나·어떤 목적으로 수집했는지 검증 가능한 형식으로 기록·공개. 집계 로그/독립 감사.

AI 특화 거부

ai-training: disallow 등 표준 디렉티브, llms.txt·옵트아웃 레지스트리. W3C 등 국제 표준화 논의와 연계.

제도: 인공지능기본법 연계

학습 데이터 출처 명시, 권리자 고지, 거부 신호 준수 확인 의무를 하위 규범·자율규제로. 다중 이해관계자 협의체로 한국형 설계.