기업 AX 워크프로세스 설계

업무 매뉴얼을
실행 가능한 규칙으로

담당자가 직접 제공한 업무 기준을 승인 가능한 정책팩으로 고정하고, PR마다 같은 기준으로 적용하는 Codex 플러그인입니다.

kakaopay-policy-review

1.000
매뉴얼 규칙 커버리지

0
환각 규칙 — 매뉴얼에 없는 규칙

1.000
재현율 — 같은 입력, 같은 결과

6 / 6
테스트용 PR 기대값 일치

카카오페이 기술블로그와 공식 페이지, 그리고 산업 리서치 한 건을 썼습니다. 내부 매뉴얼은 보지 않았고, 보지 않았다는 사실을 문서와 코드 양쪽에 명시했습니다.

기술블로그 · 카카오페이

PR을 더 느리게 만들기 위한 고민

가장 많이 참조한 자료이자 이 작업 전체의 관점이 나온 곳. AI 도입 후 코드 생산 방식은 바뀌었지만 PR을 다루는 방식은 그대로라는 문제 제기.

tech.kakaopay.com ↗

기술블로그 · 카카오페이

SDD (spec-kit) 에이전트 코딩 실전기

사내 AI 코딩·SDD 워크플로우가 이미 자리잡았음을 확인. 생성 속도가 아니라 검토 기준 적용이 병목이라는 판단의 근거.

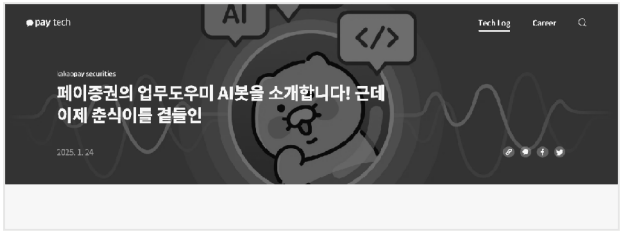
tech.kakaopay.com ↗

기업 공식 · 카카오

안전한 생성형 AI 활용을 위한 카카오페이의 노력

금융 서비스의 AI 안전·거버넌스 요구를 확인. 사람 승인 단계를 도구 구조에 박아 넣은 근거.

kakaocorp.com ↗

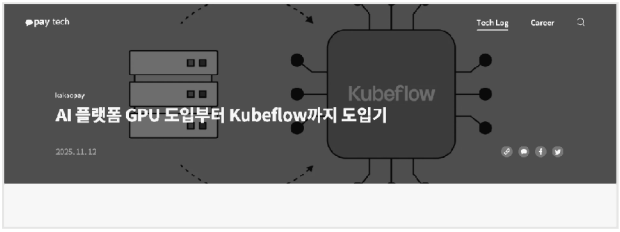


기술블로그 · 카카오페이

페이증권의 업무도우미 AI봇 (춘시리)

사내 지식이 흩어져 있고 최신성 유지가 어렵다는 점을 확인. 기준을 '문서에서 찾기'가 아니라 '정책팩으로 고정하기'로 바꾼 이유.

tech.kakaopay.com ↗



기술블로그 · 카카오페이

AI 플랫폼 GPU 도입부터 Kubeflow까지 도입기

AI 플랫폼·LLMOps 운영 제약을 확인. 새 인프라를 요구하지 않고 PR 흐름 안에서 도는 도구로 범위를 좁힌 배경.

tech.kakaopay.com ↗

산업 리서치 · GITLAB

조직이 통제 가능한 속도보다 빠르게 AI 코드를 생성한다는 조사

한 회사의 사정이 아니라 산업 전반의 문제임을 보이는 외부 근거. 관점의 일반화 가능성을 확인하는 데 사용.

ir.gitlab.com · GitLab Research 2026 ↗

이 문서는 카카오페이과 무관한 개인 작업물이며, 어떤 제휴·후원 관계도 없습니다. 캡처 2026.07.28 · 이미지는 원저작자의 것이고 출처 확인에 필요한 최소 범위만 실었습니다. 각 카드를 누르면 원문으로 이동합니다.

리서치 단계에서 이해관계자 병목 후보를 여덟 개까지 벌려 놓았고, 세운 가설은 모두 공개 자료로 확인됐습니다. 그런데 확인됐다는 것과 손대도 된다는 것은 다른 문제였습니다.

■ 확인 — 공개 자료로 뒷받침됨 ▨ 근거 부족 — 공개 자료만으로는 확인 불가 □ 폐기 — 가설 자체를 버림

가설	내용	어디서 확인했나	결과
H1	**AI가 생산 속도를 올릴수록 병목은 생성이 아니라 검토 기준 적용으로 옮겨간다**	기술블로그 「PR을 더 느리게 만들기 위한 고민」 — 코드 생산 방식은 바뀌었는데 PR을 다루는 방식은 그대로라는 문제 제기	■ 확인
H2	**사내에 이미 AI 코딩 워크플로우가 자리잡아 그 격차가 실재한다**	기술블로그 「SDD (spec-kit) 에이전트 코딩 실전기」	■ 확인
H3	**한 회사의 사정이 아니라 산업 전반의 문제다**	GitLab 2026 리서치 — 조직이 통제 가능한 속도보다 빠르게 AI 코드가 생성됨	■ 확인 · 외부 근거
H4	**금융 서비스는 사람 승인 없이 AI 산출물을 실제 행동으로 옮길 수 없다**	카카오 공식 「안전한 생성형 AI 활용을 위한 카카오페이의 노력」	■ 확인
H5	**기준을 '문서에서 찾게' 하면 최신성이 무너진다**	기술블로그 「춘시리」 — 사내 지식이 흩어져 있고 최신성 유지가 어려움	■ 확인
H6	**새 인프라를 요구하는 도구는 실제로 도입되지 않는다**	기술블로그 「AI 플랫폼 GPU 도입부터 Kubeflow까지」 — 운영 제약 확인	■ 확인

되돌리기 · 범위 회귀

제일 먼저 정한 것은 만들 것이 아니라 하지 않을 일이었다

리서치에서 확인된 병목은 여덟 개였습니다 — 개발자 리뷰, AI 플랫폼 표준화, GPU 운영, 사내 지식, FDS, 보험 문서, 결제 에이전트 경계, 서비스 탐색. 가설은 전부 공개 자료로 확인됐지만, 확인됐다는 것과 만들어도 된다는 것은 다릅니다. 그래서 하지 않을 일을 먼저 못 박았습니다. 내부 매뉴얼을 안다고 주장하지 않는다. 리뷰 도구를 하나 더 추가하지 않는다. 새 인프라를 요구하지 않는다. 세 개를 빼고 나니 PR 흐름 안에서 도는 얇은 레이어만 남았습니다.

이해관계자 병목 8개 → 가설은 전부 확인 → **범위 회귀** → 하지 않을 일 3개 확정 → **PR 흐름 안의 레이어** → 테스트 6건으로 재검증

> 최종 문제 정의 — 새 리뷰 도구를 하나 더 추가하는 것이 아니라, **이미 있는 기준을 실행 가능하게 만든다.**

금융 서비스에서는 산출물이 내부 업무 기준 · 보안 기준 · 규제 리스크 · 승인 절차를 통과해야 합니다.
그런데 그 기준은 문서와 담당자 경험에 흩어져 있습니다.

AI 도입 이전

생산과 검토가 비슷했다

코드 생산 속도와 사람의 리뷰 용량이 대체로 균형을 이뤘습니다.

현재

생산이 검토를 넘어섰다

생산 속도가 리뷰 용량을 넘어섰고, 그 격차가 병목입니다. 이 플러그인이 서는 자리입니다.

기준이 있는 곳

문서와 사람의 머릿속

보안·개인정보, 결제 도메인, AI 안전, 운영 기준이 각각 다른 곳에 흩어져 있습니다.

이 도구를 쓰는 사람 — 두 역할이 한 흐름에서 만난다

기준 제공자	**보안·개인정보, 결제 도메인, AI 안전, 운영 담당자.** 자기 매뉴얼을 넣고 규칙 후보를 승인합니다.
PR 작성자 · 리뷰어	고정된 정책팩 기준으로 검토를 받습니다. 통과 항목은 보이지 않고 예외와 근거 부족만 올라옵니다.
쓰는 구간	AI가 만든 변경을 포함해, PR을 열고 머지하기 전 구간.

플러그인은 무엇으로 이루어져 있나

쓰는 사람

기준 제공자

보안·개인정보, 결제, AI 안전, 운영 담당자. 매뉴얼을 넣고 규칙을 승인합니다.

PR 작성자 · 리뷰어

승인된 기준으로 검토를 받습니다.

Codex 앱 대화창

새 리뷰 도구를 띄우지 않고, PR을 다루던 흐름 안에서 씁니다.

플러그인 — 이 프로젝트가 만든 것

플러그인 진입점

`src/.codex-plugin/plugin.json`

Codex가 이 폴더를 플러그인으로 알아보게 하는 파일.

사용 설명서 (스킬)

`skills/policy-manual-review/SKILL.md`

매뉴얼을 규칙으로 바꿀 때 지켜야 할 경계를 적어 둔 문서.

정책 스크립트

`scripts/policy_harness.py`

매뉴얼을 규칙으로 바꾸고, PR을 검토하고, 결과를 평가하는 세 가지 일을 합니다.

정책팩 · 테스트용 PR

`evals/`

승인된 규칙 묶음과, 기대 결과를 미리 적어 둔 PR 6건.

드나드는 데이터

들어오는 것

담당자가 준 업무 매뉴얼·체크리스트·PR 템플릿, 그리고 검토할 PR의 변경 내역과 본문.

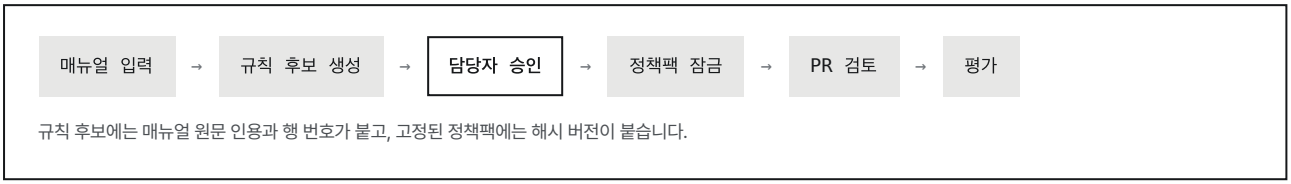
나가는 것

예외와 근거 부족만 추린 리뷰 리포트, 그리고 규칙이 기대대로 작동했는지 확인하는 평가 리포트.

규칙은 담당자가 승인하기 전까지 PR에 닿지 않습니다. 승인 단계가 흐름 중간이 아니라 구조에 박혀 있습니다.

사람 승인 단계가 구조에 박혀 있다

AI가 스스로 정책을 만들어 적용할 수 없습니다. 승인되지 않은 규칙은 PR에 닿지 않습니다. 이것은 나중에 덧붙인 안전장치가 아니라, 내부 매뉴얼을 안다고 주장하지 않기로 한 결정에서 나온 결과입니다.



담당자가 확정하기 전에는 어떤 규칙도 PR에 적용되지 않는다

판정	의미	리뷰어 화면
사람이 판단 `human_review`	예외 트리거가 걸림 — 자동 차단이 아니라 담당 조직으로 넘김	보임
근거 부족 `evidence_missing`	규칙은 적용되나 필수 근거가 PR에 없음	보임
통과 `pass`	규칙 적용, 근거 있음	기계 리포트에만
해당 없음 `not_applicable`	이 PR에 해당하지 않는 규칙	기계 리포트에만

지적 하나의 형태

심각도 / 담당 조직 · 매칭 트리거 · **빠진 근거 필드** · 매뉴얼 출처(행 번호와 원문 인용) · 다음 행동 · 정책 버전 해시가 한 덩어리로 붙습니다.

설계하면서 내린 판단

매뉴얼에 없는 규칙은 안 만든다

모든 규칙은 원문 인용과 행 번호에 묶입니다. 평가 항목에 환각 규칙 수를 넣어 이 경계를 수치로 지킵니다.

최종 판단은 사람에게

예외 트리거가 걸리면 자동 차단이 아니라 담당 조직으로 넘깁니다.

정책을 해시로 고정

정책팩에 해시 버전을 붙여, 어떤 기준으로 검토된 PR인지 나중에도 확인할 수 있게 했습니다.

로그를 안전하게 남기는 방식까지

해커톤은 AI 대화 로그 제출을 요구했습니다. 원문 전체 복사 폴백을 끄고, 안전한 메타데이터와 원본 라인 수·SHA-256 해시만 남기며, 요약만 덧붙여 쌓는 방식으로 로거를 교체했습니다.

평가 항목	결과	의미
매뉴얼 규칙 커버리지	1.000	매뉴얼 조항이 빠짐없이 규칙으로 변환됨
환각 규칙 수	0	매뉴얼에 없는 규칙을 만들지 않음
규칙 상태 정확도	1.000	위반 · 근거부족 · 사람확인을 기대값대로 구분
오탐 수	0	정상 변경 PR에 불필요한 지적 없음
재현율	1.000	같은 입력에 같은 결과
기대값 대조	6 / 6	정상 변경 · PII 로그 · 환불 근거부족 · 환불 근거완비 · AI 평가 근거부족 · 롤백 미비

기대할 수 있는 것 — 검증된 성과가 아니라 이 도구가 노리는 지점

리뷰 부하 축소

통과 항목 대신 예외와 근거 부족만 리뷰어에게 올라간다.

추적 가능한 근거

매뉴얼 조항과 PR 근거가 연결되고, 정책 버전이 해시로 고정된다.

기준 편차 축소

같은 입력에 같은 결과가 나와 사람 간 판단 차이가 줄어든다.

정책 권한의 위치

사람 승인 단계가 구조에 있어 AI가 정책을 스스로 만들 수 없다.

주장하지 않은 것

내부 매뉴얼을 안다고 하지 않았다

데모 정책팩은 예선 검증용으로 직접 작성한 4개 규칙이며, 문서와 코드 양쪽에 그렇게 명시했습니다.

사람 리뷰를 대체하지 않는다

예외는 자동 차단이 아니라 담당 조직에 넘깁니다.

정확도는 데모 조건 안의 수치다

1.000은 테스트용 PR 6건과 4개 규칙 범위에서의 결과입니다.

운영 배포를 하지 않았다

로컬 Codex 앱에서 플러그인 인식과 skill 동작까지 검증한 상태입니다.

설계하면서 느낀 점

제일 먼저 정한 건 만들 것이 아니라 하지 않을 일이었습니다. 리서치로 확인된 병목이 여덟 개였는데, 확인됐다는 것과 손대도 된다는 것은 다르다는 걸 그때 알았습니다. 내부 매뉴얼을 안다고 주장하지 않기로 정하자 도구의 모양이 저절로 정해졌습니다. 규칙을 사람이 승인하기 전에는 PR에 닿지 않는 구조가 된 것은, 안전을 위해 덧붙인 장치가 아니라 그 결정에서 자동으로 나온 결과였습니다.