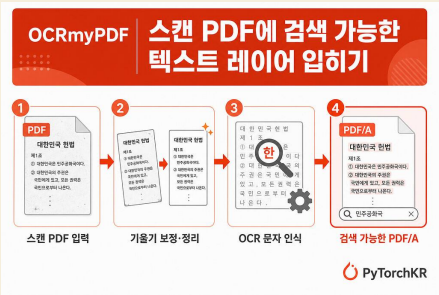
THE WEEKLY DAEMON

Vol.17 2026.7/2 - 7/9 WEEKLY TECH & AI DIGEST 25 Articles This Week

OCRmyPDF — 스캔 PDF에 검색 가능한 텍스트 레이어 자동 추가



스캐너로 읽은 PDF는 이미지 기반이라 텍스트 검색과 복사가 불가능한 문제가 있다. OCRmyPDF는 Tesseract OCR 엔진을 활용하여 원본 이미지는 유지하면서 그 아래에 인식된 텍스트 레이어를 무손실로 삽입한다. 100개 이상의 언어를 지원하고, 기울어진 페이지 보정, 이미지 최적화, PDF/A 표준 출력 등 실무용 기능을 제공하며, 모든 처리가 로컬에서 이루어져 개인정보 보호를 보장한다.

> 핵심 기능: 다국어 100개 이상 지원, 원본 이미지 해상도 유지, 텍스트 인식 후 복사·붙여넣기 오류 제거, 기울기 보정·이미지 최적화로 입력 파일보다 작은 결과물 생성.

LINK discuss.pytorch.kr/t/ocrmypdf-pdf/11047

pxpipe — Claude 코드 요청 토큰을 이미지로 변환해 59~70% 비용 절감

Claude Code 사용 시 시스템 프롬프트, 도구 문서, 긴 히스토리 같은 대량의 텍스트 컨텍스트가 입력 토큰을 크게 증가시키는 문제를 해결한다. pxpipe는 이런 밀집 콘텐츠를 PNG 이미지로 렌더링하는 로컬 프록시로, 이미지 토큰 비용이 픽셀 크기에만 의존한다는 특성을 활용한다. 실제 Claude Code 트래픽에서 텍스트는 토큰당 약 1자를 담지만 이미지는 약 3.1자를 담아, Fable 기준 엔드투엔드 청구액을 약 59~70% 낮춘다.

> **핵심 성과:** 48k 문자의 시스템 프롬프트와 도구 문서를 약 25k 텍스트 토큰 대신 약 2.7k 이미지 토큰으로 압축하며, 실제 세션 비용을 6.06달러로 낮춤(텍스트 기준 42.21달러).

LINK news.hada.io/topic

DRE 측정 및 감소 — LLM의 테이블 데이터 참조 오류 체계적 평가

대형 언어 모델이 테이블 작업에서 테이블 구조를 이해하면서도 값을 잘못 인용하거나 생략하는 데이터 참조 오류(DRE)를 발생시키는 문제를 해결하는 연구다. 크리틱 기반 필터링과 거부 샘플링을 통해 답변 정확성을 최대 12.0% 향상시키며, 4B 파라미터 크리틱 모델은 분포 내외 DRE 탐지에서 78.2% F1 스코어를 달성하여 대규모 모델의 추론을 효과적으로 지원한다.

> **핵심 성과:** 1.7B~20B 파라미터 모든 모델에서 DRE 발생 확인, 크리틱 기반 접근으로 답변 정확성 최대 12.0% 개선, 가벼운 4B 파라미터 크리틱 모델이 78.2% F1 스코어로 분포 내외 오류 탐지 및 대형 모델 추론 보조

LINK arxiv.org/abs/2606.32029

SGR-BIM — 그래프 기반 의미론적 추론으로 건설 규정 준수 자동화

건설정보모델링(BIM)에서 기하학적 규정 준수 확인은 고수준의 규제 논리와 구조화된 IFC 데이터 간 의미론적 불일치로 인해 자동화가 어려운 과제였다. SGR-BIM은 교차모달 지식그래프를 동적으로 구축하여 사용자 의도, 규제 의미론, BIM 기하학을 정렬하고 경직된 규칙 템플릿 없이도 해석 가능한 추론을 수행한다. 이를 통해 다단계 추론 체인 순회와 건물 엔티티 간 잠재적 공간 의존성 해결을 가능하게 한다.

> **핵심 성과:** 화재 안전 코드 679개 전문가 검증 쿼리에서 84.3% 정확도 달성, 기존 단일 에이전트 대비 8.6% 개선

LINK arxiv.org/abs/2606.12065

Grok 4.5 — SpaceXAI가 Cursor와 공동 개발한 범용 AI 모델 공개

SpaceXAI가 Cursor와 협력하여 개발한 Grok 4.5를 공개했다. 이 모델은 기존의 코딩 전용 모델에서 벗어나 소프트웨어 개발, 데이터 과학, 금융, 법률 업무 등 다양한 분야를 지원하는 범용 모델이다. MoE 구조를 채택하여 필요한 전문가만 활성화하여 효율성을 높였으며, Claude Opus와 GPT 5.5에 필적하는 성능을 제공하면서도 비용은 10분의 1 수준으로 낮췄다.

> **핵심 성과:** 1.5조 파라미터 V9 기반 모델로 Opus급 성능을 구현하면서 경쟁 모델 대비 10분의 1 수준의 비용으로 제공. 수만 장의 NVIDIA GB300 GPU를 활용한 대규모 학습으로 코딩부터 지식 작업까지 다양한 작업에서 기존 모델을 초과하는 성능 달성.

LINK x.ai/news/grok-4-5

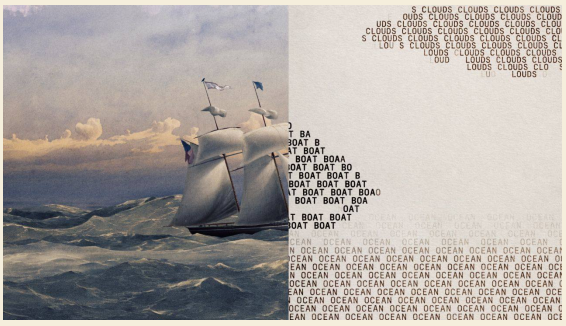
Claude: 모델 업그레이드 vs Effort 증가, 언제 어떻게?

Claude 코드 생성 결과가 만족스럽지 않을 때 모델을 올릴지 effort를 올릴지 혼동하는 사용자들을 위해 Claude 팀이 공식 가이드라인을 제시했다. Effort는 단순한 '더 오래 생각하기'가 아니라 파일 읽기, 도구 활용, 테스트 실행 등 전체 작업 범위를 결정하는 값이며, 모델은 기본 능력의 상한을 결정하는 완전히 다른 축이다. 실패 원인을 먼저 분석해 지식 부족으로 인한 오류인지 불충분한 실행으로 인한 오류인지 판단한 후 적절한 파라미터를 조정해야 한다.

> **핵심 기여:** Effort는 컨텍스트 크기, 도구 사용 범위, 테스트 횟수를 포함한 작업 범위를 제어하는 값이며, 모델은 기본 능력 자체를 결정하는 독립적인 축이라는 명확한 구분 제시. 사용자가 실패 원인을 올바르게 진단하면 모델과 effort 파라미터를 효과적으로 최적화할 수 있다.

LINK www.threads.com/@info__sum/post/DagwO...

Anthropic: Claude 내부 '생각의 자리' J-공간 발견



대형 언어모델 Claude가 외부에 드러내지 않은 내부 사고 과정을 가지고 있다는 문제를 해결하기 위해 Anthropic이 신경망 내부 패턴을 분석했다. J-공간이라 명명된 이 영역에서 Claude는 사람이 의식적으로 접근 가능한 사고와 유사한 방식으로 중간 생각들을 처리하며, 이는 학습 과정에서 자발적으로 생성되었다. 연구진은 이 영역의 내용을 조작해 Claude의 응답을 제어할 수 있음을 입증했다.

> **핵심 성과:** Claude의 내부 표현을 조작하여 명시적 지시 없이도 응답 변경 가능(예: 거미 관련 문제에서 '거미'를 '개미'로 치환하면 답이 8에서 6으로 변경됨), AI 모델에서 의식적 접근성과 유사한 구조의 자발적 출현 확인.

LINK www.anthropic.com/research/global-wor...

Self-Harness — LLM 에이전트의 자동 하네스 최적화 루프

LLM 에이전트의 성능은 모델 자체뿐 아니라 시스템 프롬프트, 도구 규칙, 실행 정책 등으로 구성된 하네스에 의해 결정되는데, 이 하네스 설계가 전적으로 인간의 몫이었다는 한계가 있다. Self-Harness는 에이전트가 자신의 실패 트레이스를 분석해 하네스를 스스로 개선하는 자동화 루프를 제안한다. 실패 패턴 추출, 수정안 생성, 회귀 테스트를 통한 검증 단계를 거쳐 외부 모델이나 인간 개입 없이 독립적으로 성능을 향상시킨다.

> **핵심 기여:** 모델 가중치 변경 없이 하네스 자동 최적화만으로 SWE-bench 성능을 20%에서 50%로 향상시켰으며, 에이전트가 자신의 실패 기록으로부터 하네스 수정안을 스스로 제안하고 검증하는 자기개선 루프를 구현했다.

LINK lilianweng.github.io/posts/2026-07-04...

NVIDIA ICML 2026: LLM 메모리 용량 3.6비트/파라미터 규명

LLM이 학습 데이터를 얼마나 암기할 수 있는지 측정하는 것이 중요한 과제였다. NVIDIA의 최신 ICML 논문은 파라미터당 약 3.6비트의 메모리 용량을 정량화하여, 단순 암기와 진정한 일반화 능력을 명확히 구분했다. 이 연구는 향후 모델 크기 결정 및 데이터 프라이버시 정책 수립 시 과학적 기준점으로 활용될 수 있다.

> **핵심 성과:** LLM의 메모리 용량을 파라미터당 3.6비트로 정량화하여 암기 능력과 일반화 능력의 경계를 명확히 규정, 모델 스케일링 및 데이터 프라이버시 의사결정의 객관적 기준 제공

LINK blogs.nvidia.com/blog/open-models-icm...

BIM Information Extraction Through LLM-based Adaptive Exploration — LLM 에이전트 기반 적응형 탐색

BIM 모델에서 정보 추출 시 고정된 쿼리 방식의 한계를 해결하는 논문. 각 프로젝트마다 구조가 다른 BIM의 이질성 문제를 LLM 에이전트가 런타임 중에 모델 구조를 직접 파악하며 코드를 반복 실행하는 적응형 탐색 패러다임으로 극복한다. 37개 IFC 모델과 1,027개 질의로 구성된 벤치마크에서 정적 쿼리 생성 방식을 모든 조합에서 능가했다.

> **핵심 성과:** 21개 프로젝트 37개 IFC 모델 1,027개 질의 벤치마크에서 적응형 탐색이 정적 쿼리 생성을 모든 조합에서 초과 성능 달성, BIM 활용의 병목이 쿼리 설계에서 모델 탐색 방식으로 전환.

LINK arxiv.org/abs/2605.01698

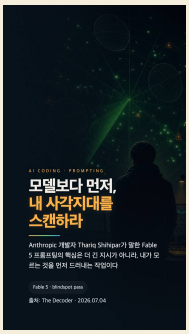
SkillWeaver — AI 에이전트의 다중 스킬 조합 문제를 분해-검색-조합으로 해결

LLM 에이전트에 수많은 도구를 한 번에 제공하면 오히려 성능이 저하되는 문제를 해결하는 프레임워크. SkillWeaver는 복잡한 사용자 쿼리를 원자적 부작업으로 분해하고, 각 부작업에 맞는 스킬을 검색한 후, 의존성을 고려한 DAG 계획으로 실행 전 완벽한 계획을 수립한다. 2,209개의 실제 MCP 서버 스킬을 대상으로 평가하는 CompSkillBench 벤치마크를 제시한다.

> **핵심 기여:** 복합 작업 환경에서 최고 모델도 완전 해결률 3% 수준인 multi-week 과제를 분해-검색-조합 전략으로 성능 향상, 300개 조합 쿼리를 포함한 벤치마크 제공으로 에이전트 스킬 라우팅 평가 기준 수립.

LINK arxiv.org/abs/2606.18051

Fable 5: 모델 성능보다 명확한 프롬프팅이 결과를 결정



앤트로픽의 최신 모델 Fable 5는 거의 모든 벤치마크에서 최고 수준의 성능을 기록했으나, 창시자 Thariq는 모델의 강력함이 증대될수록 작업 품질의 병목이 모델 자체가 아니라 사용자의 프롬프팅 명확성으로 이동한다고 지적한다. 프롬프트(map)와 실제 작업 환경(territory) 사이의 간극에서 발생하는 미지의 영역을 줄이는 것이 결과 품질을 좌우하며, 이는 Known Knowns, Known Unknowns, Unknown Knowns, Unknown Unknowns 4단계 분석을 통해 체계적으로 관리할 수 있다는 실전 가이드를 제시했다.

핵심 기여: 초강력 AI 모델 시대에 작업 품질 향상의 핵심이 프롬프팅 정확도와 컨텍스트 완성도로 이동했음을 규명하고, 사용자가 명확하지 않은 요구사항을 체계적으로 파악하고 전달하는 4단계 프레임워크 제시.

LINK x.com/trq212/status/2073100352921215386

EdgeBench — AI 에이전트의 장기 학습 능력을 측정하는 벤치마크

기존 벤치마크는 AI 모델의 현재 지식을 평가하는 스냅샷에 불과하다는 문제를 해결하기 위해 바이트댄스 Seed Edge 팀이 EdgeBench를 공개했다. 이 벤치마크는 AI 에이전트에게 중력파 분석, CPU 설계, 수학 정리 증명 등 실무 전문가가 며칠 소요하는 복잡한 과제 134개를 최대 72시간 동안 수행하게 하며, 피드백을 받아 반복적으로 개선하는 과정을 통해 진정한 학습 곡선을 측정한다. 134개 과제 평균 결과가 R^2=0.998의 완벽한 S자 곡선을 이루며 물리 법칙 수준의 확정성을 보여준다.

핵심 성과: 평균 134개 과제에서 오차율 R^2=0.998의 매끄러운 확장 법칙 도출, 기존 정적 벤치마크의 한계를 극복하고 AI 에이전트의 실시간 학습 궤적과 차세대 스케일링 방향을 정량적으로 측정 가능하게 함.

LINK www.threads.com/@choi.openai/post/DaT...

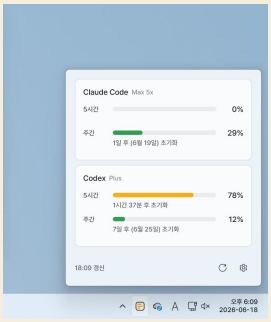
Graphify — AI 코딩 비용 71.5배 절감하는 지식그래프 기반 컨텍스트 압축

AI 코딩 보조 도구 사용 시 실제 비용이 모델 성능이 아닌 반복되는 컨텍스트 입력에서 발생하는 문제를 해결한다. Graphify는 코드, SQL 스키마, 문서, PDF, 이미지를 지식그래프로 변환하여 Claude Code에서 직접 활용할 수 있도록 한다. 쿼리당 필요한 토큰을 71.5배 감소시켜 AI 코딩의 경제성을 크게 개선하며, 컨텍스트 압축을 통해 대규모 프로젝트에서도 효율적인 코드 생성을 가능하게 한다.

핵심 성과: 기존 대비 쿼리당 토큰 사용량 71.5배 감소, GitHub 79,966 스타 획득. Claude Code와 Codex, OpenCode, Cursor, Gemini CLI 등 주요 AI 코딩 어시스턴트 지원.

LINK github.com/Graphify-Labs/graphify

CodeIsland — 맥북 노치에서 AI 코딩 에이전트 상태를 실시간 모니터링

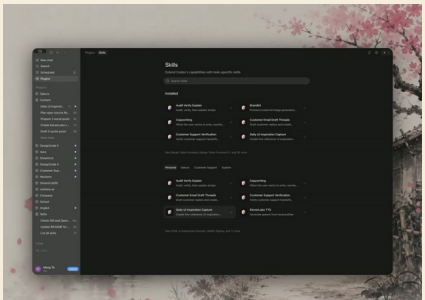


여러 AI 코딩 에이전트를 동시에 실행할 때 각 세션의 상태를 확인하기 위해 창을 오가는 비효율이 발생하는 문제를 해결한다. CodeIsland는 MacBook 노치 영역에 상주하면서 Claude Code, Codex, Gemini CLI 등 13종의 AI 도구에서 발생하는 세션 시작, 권한 요청, 작업 완료 등의 이벤트를 유닉스 소켓 기반 IPC로 수집하여 컴팩트한 픽셀 아트 패널에 실시간 표시한다. 패널에서 직접 권한 승인 및 거부가 가능하며, 작업 중인 앱을 떠나지 않고 에이전트의 질문에 응답할 수 있다.

핵심 기여: 13종의 주요 AI 코딩 도구에 자동 훅 설치하여 이벤트 수집, Swift 기반 오픈소스로 macOS 14 이상에서 원클릭 점프 기능 및 에이전트별 픽셀 아트 마스코트 지원

LINK discuss.pytorch.kr/t/codeisland-ai-ma...

Agent Skills — Meng To의 AI 코딩 에이전트용 75개 디자인 스킬 오픈소스 공개

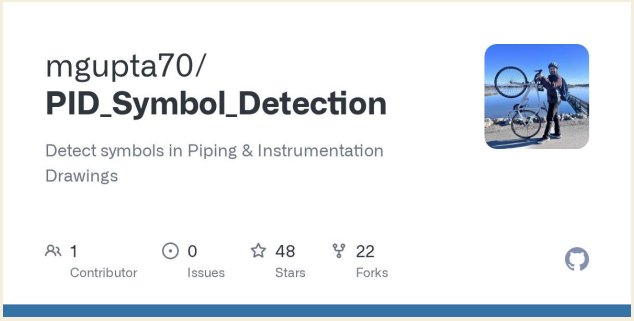


웹디자인과 프론트엔드 개발 작업에서 AI 에이전트의 활용도가 낮은 문제를 해결하기 위해 Design+Code 창업자 Meng To가 Agent Skills 라이브러리를 오픈소스로 공개했다. Codex, Claude Code, Cursor 등의 AI 코딩 에이전트에서 즉시 사용 가능한 75개의 스킬을 제공하며, 화면 녹화를 Fable 5용 프롬프트로 변환하거나 기존 HTML 페이지를 상호작용 프롬프트로 추출하는 기능으로 웹디자인, 랜딩페이지, 모션, WebGL, UI 제작 작업을 대폭 간소화한다.

핵심 기여: Video to Super Prompt, HTML to Interaction Prompts 등 고도화된 워크플로우 제공으로 AI 에이전트 기반 디자인-개발 작업을 자동화하며, DESIGN.md 기반 디자인 시스템 분석으로 일관된 시각 언어 유지를 가능하게 한다.

LINK github.com/MengTo/Skills

PIDSymbolDetection: 범례 한 장으로 배관도 기호 자동 인식

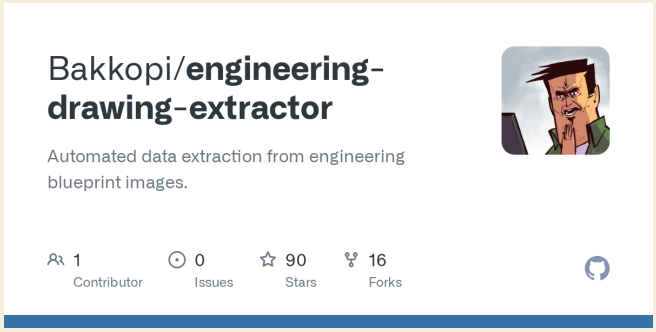


도면 인식 프로젝트에서 새 기호마다 수백 장의 라벨링 데이터가 필요한 문제를 해결하는 오픈소스. 2단계 파이프라인으로 첫 단계는 클래스무관 YOLO 탐지기가 기호 영역을 추출하고, 두 번째 단계는 범례 이미지 한 장으로 few-shot 분류를 수행하여 실제 기호명을 결정한다. 라벨링 비용이 기호 개발 속도를 못 따라가는 P&ID 도면 인식의 병목을 정면으로 해결한다.

핵심 기여: 범례 이미지 1장만으로 새 기호 추가 가능한 few-shot 분류 방식 도입으로 라벨링 총량 제거, MIT 라이선스 오픈소스 프로젝트(별 48개, Python 100%, 87커밋)로 초기 단계이지만 파일럿 검증 수준에서 실무 활용 가능성 제시

LINK github.com/mgupta70/PID_Symbol_Detection

engineering-drawing-extractor: 도면 표제란 자동 추출 도구

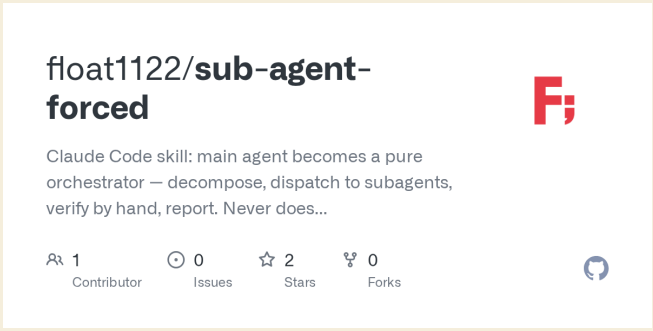


엔지니어링 도면 수백 장에서 표제란의 도면번호, 작성자, 제목 등을 수동으로 정리하는 반복적인 작업이 문제다. engineering-drawing-extractor는 도면 이미지를 입력받아 도면 영역과 표제란 표를 분리한 후 OCR을 통해 핵심 메타데이터를 자동 추출하는 경량 파이썬 도구다. 테두리와 불필요한 선, 주석을 제거해 인식 정확도를 높인 실용적 파이프라인으로, 이미 정리된 도면에서 메타데이터를 추출해 목록화하는 데 특화되어 있다.

> **핵심 기여:** 불필요한 요소 제거를 통한 OCR 정확도 향상으로 도면 메타데이터 추출 자동화를 실현했으며, 규모가 작아 현장 도면 양식에 맞춘 커스터마이징이 용이하다.

LINK github.com/Bakkopi/engineering-drawin...

Hermes Agent Kanban — 다중 에이전트 협업 워크플로우 관리 시스템

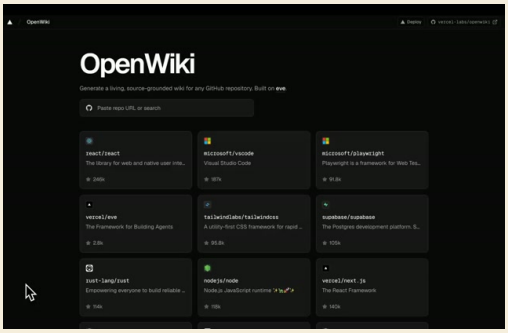


여러 에이전트로 작업을 분담할 때 각 에이전트의 작업 상태를 추적하기 어려운 문제를 해결하는 Hermes의 Kanban 기능. 메인 에이전트를 오케스트레이터로 설정하고 AGENTS.md를 통해 칸반 워크플로우를 명시하며, 이슈 트리아지에서 작업 분배, 자동 QA 검수, 리뷰까지의 전체 파이프라인을 티켓 단위로 관리한다. QA 실패 시 피드백 루프가 자동으로 작동하고, 모든 작업 로그가 티켓에 누적되어 감사 추적이 용이하다.

> **핵심 기여:** 에이전트 간 작업 분배와 진행 상태를 시각적으로 추적하며, 티켓 기반의 완전한 감사 로그로 각 작업의 전체 컨텍스트와 피드백 히스토리를 보존한다.

LINK github.com/float1122/sub-agent-forced

OpenWiki — GitHub 코드 변경 시 자동 갱신되는 오픈소스 위키 생성기

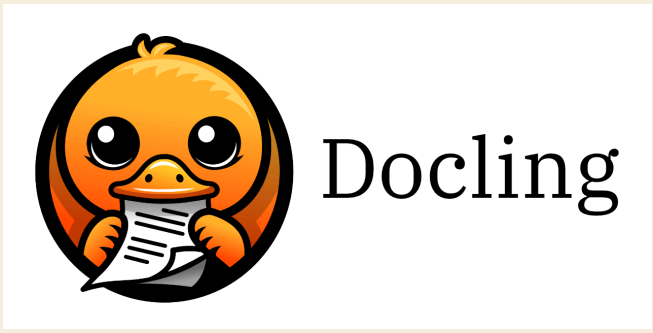


코드가 수정되어도 문서는 낡은 채로 방치되는 문제를 해결하는 도구. OpenWiki는 GitHub 저장소를 입력받아 코드 변경 시마다 문서 사이트를 자동으로 재생성하며, 사람이 수동으로 유지보수하는 방식을 벗어나 빌드 프로세스처럼 소스 코드에서 파생된 문서를 구조적으로 유지한다. Vercel의 eve AI 에이전트와 Neon, Blob을 활용한 백그라운드 작업으로 신선한 문서를 자동 생성한다.

> **핵심 기여:** 에이전트가 파생물을 생성해 신선하게 유지하는 앱의 표준형을 구현했으며, 코드 분석과 위키 생성을 백그라운드 작업으로 처리하여 문서 노후화 문제를 구조적으로 차단한다.

LINK github.com/sodam-ai/wikimate

Docling — 논문 기준 최고 성능의 PDF 파싱 솔루션

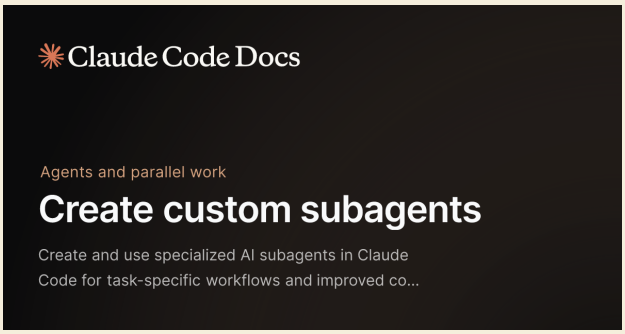


PDF 파싱 도구들을 비교 평가한 결과 Docling이 논문 기준으로 가장 우수한 성능을 보여준다. 처리 시간은 다소 소요되지만 문서 구조를 정확하게 복원하여 생성 AI 에코시스템과의 연동에 최적화되어 있다. 다양한 문서 형식 지원, 고급 PDF 이해 기능, 통합 문서 표현 형식 등을 통해 AI 기반 데이터 추출의 신뢰성을 확보한다.

> **핵심 기여:** PDF 외에도 DOCX, PPTX, XLSX, HTML, EPUB 등 다양한 형식 지원하며, 페이지 레이아웃, 읽기 순서, 표 구조, 수식, 이미지 분류 등 고급 PDF 이해 기능으로 0.928의 표 정확도 달성.

LINK github.com/docling-project/docling

Claude Code — 고가 모델은 계획만, 저가 모델이 코드 작성하는 하이브리드 프롬프트

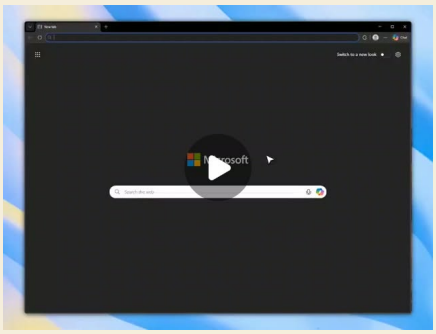


Claude Fable 5는 비싸고 강력하지만 실제 코드 작성에는 비효율적이라는 문제를 해결하기 위해, 고안된 방식이다. 터미널 명령 한 줄로 Fable 5를 계획 역할에만 제한하고, 저가 모델인 Claude Sonnet 5에게 실제 코드 작성을 위임하는 방식이다. 환경변수와 시스템 프롬프트를 통해 서브에이전트의 모델을 고정하고 역할을 명확히 하며, 추가 도구나 설정 파일 없이 Claude Code의 기본 기능만으로 구현 가능하다.

> **핵심 성과:** 설정 파일 없이 터미널 한 줄 명령으로 다중 모델 조율을 실현하며, 고가 모델의 사용률을 최소화하면서도 코드 생성 품질을 유지하는 비용 최적화 프롬프팅 기법 제시

LINK code.claude.com/docs/en/sub-agents

Astryx — Meta의 AI 에이전트 대응 디자인 시스템 공개

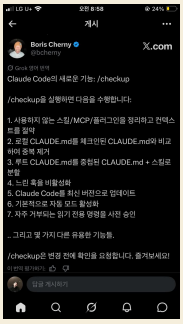


Meta가 8년간 내부에서만 사용해온 디자인 시스템 Astryx를 공개했다. 인간 개발자뿐 아니라 AI 코딩 에이전트까지 고려해 설계된 이 시스템은 13,000개 이상의 내부 앱에서 검증되었으며, 150개 이상의 UI 컴포넌트, 테마, 브랜드 스킨을 제공한다. StyleX 기반이면서도 독립적으로 사용 가능하고, AI 에이전트가 쉽게 활용할 수 있도록 문서와 API가 구성되어 있어 AI 기반 프런트엔드 개발 환경에 적합하다.

> **핵심 성과:** 13,000개 이상의 사내 앱에서 검증된 150개 이상의 UI 컴포넌트 제공, AI 에이전트 친화적 API 및 문서 구조로 인간-AI 협업 개발 환경 지원

LINK www.threads.com/@choi.openai/post/DaS...

Claude Checkup — 불필요한 스킬과 플러그인 정리 기능 출시

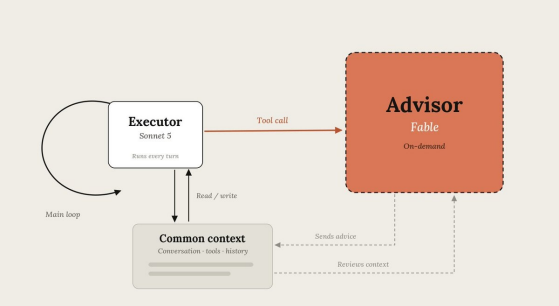


Claude Code에서 프로젝트에 불필요하게 쌓인 스킬과 플러그인을 체계적으로 정리하기 어려운 문제를 해결하기 위해 /checkup 기능을 출시했다. Claude.md를 기준점으로 삼아 중구난방하게 구성된 개발 환경을 한 번에 점검하고 정리할 수 있게 지원하여 프로젝트 유지보수의 효율성을 높인다.

> 핵심 기능: Claude Code 내 /checkup 명령으로 프로젝트의 불필요한 스킬, 플러그인, 기술 스택을 자동 진단하고 정리 제안

LINK www.threads.com/@unclejobs.ai/post/Da...

Claude Advisor — 멀티모델 라우팅을 설정 한 줄로 구현



AI 에이전트 운영 시 모든 작업을 고성능 모델에 맡기면 비용이 과다하게 발생하는 문제를 해결하기 위해 Anthropic이 Claude Code에 내장한 어드바이저 기능. 저비용 모델이 주요 작업을 실행하다가 어려운 판단이 필요할 때만 고성능 모델에 조언을 요청하는 멀티모델 라우팅을 한 줄 명령어로 활성화할 수 있다. 호출은 Anthropic 서버가 처리하므로 별도 인프라가 필요 없고, 메인 모델의 프롬프트 캐시도 유지된다.

> **핵심 기여:** 수동으로 관리해야 하던 멀티모델 라우팅 규칙을 설정 한 줄로 자동화하고, 어드바이저는 400~700토큰의 짧은 지침만 생성하여 주요 비용을 차지하는 실행 작업은 저비용 모델에 맡기는 구조 실현.

LINK www.threads.com/@unclejobs.ai/post/Da...