A DAILY PERSPECTIVE ON AI

AP *News*

필요한 정보만, 실제 적용까지 확인하기

AI 도구가 많아질수록 무엇을 불러오고 어디서 실행하며 어떤 정보를 옮길지 정하는 일이 중요해집니다. 오늘은 필요한 도구의 선택, 요청 분배, 설정 적용과 맥락 이전을 살펴봅니다. 기능 소개·절차 안내·고객 사례를 구분해 읽습니다.



AI 제작 · 편집 일러스트 · 실제 제품 화면 아님

THE DAILY BRIEF · 5 STORIES

전날인 2026년 11월 22일 이후 공식 발표를 우선 조사했으나 이전 호와 같은 이슈를 제외한 원문이 부족해 선정 범위를 7일로 넓혔습니다. 실제 선정 원문은 11월 19~20일의 Google 3건·AWS 2건입니다. 대화 이전 글은 새 출시가 아닌 재안내, 기업 활용 글은 고객 사례로 구분했습니다. AI 작성 및 원문 확인.



AI 제작 · 편집 일러스트 · 실제 제품 화면 아님

01 · AI NEWS

Gemini Tool Retrieval, 필요한 도구 설명만 그때 불러오기

Google Blog · 발표 2026-11-19

Google은 도구의 매개변수 정의를 필요할 때 불러오는 기능을 Interactions API 미리보기로 발표했습니다.

Google은 11월 19일 Gemini API의 Tool Retrieval을 소개했습니다. 모델은 미리 불러오지 않은 도구의 이름·설명은 보고, 필요할 때 검색해 전체 매개변수 정의를 가져온 뒤 호출을 요청합니다. 실제 기능 실행과 결과 반환은 기존처럼 애플리케이션 코드가 맡습니다. 서버 검색과 원격 MCP를 지원하며 자체 검색 방식을 쓰면 검색마다 추가 왕복이 발생합니다. Gemini 3.6 이상이 대상입니다. 기능 자체의 별도 요금은 없지만 모델의 입출력 토큰은 과금되며, 불러온 정의도 이후 대화의 입력으로 계산되므로 실제 절감은 사용 패턴에 달려 있습니다.

AP의 관점 · AP의 웹·AI 교육에 참고한다면, 가상 도구 목록에서 이름·설명과 상세 입력 정의를 분리해 비교하는 과제를 제안합니다. 도구를 더 많이 연결하는 것보다 필요한 도구를 정확히 고르고 실행 권한을 따로 확인하는 기준을 설명할 수 있습니다.

공식 원문 읽기 ↗



AI 제작 · 편집 일러스트 · 실제 제품 화면 아님

02 · AI NEWS

HyperPod Global Gateway, 여러 클러스터를 한 진입점으로

AWS Blog · 발표 2026-11-20

AWS는 리전·클러스터 사이에 요청을 나누는 관리형 추론 게이트웨이를 소개했습니다.

AWS는 11월 20일 SageMaker HyperPod Global Inference Gateway를 발표했습니다. 등록한 클러스터의 여유 메모리·대기열·리전 거리·오류율을 바탕으로 요청을 보내고, 클러스터 내부의 기존 게이트웨이가 모델 작업을 고릅니다. SigV4·IAM 인증과 PrivateLink 연결을 사용합니다. 실패 요청은 스트리밍이 시작되기 전이면 다른 클러스터로 한 번 재시도하고, 정상 클러스터가 없으면 503을 반환합니다. 싱가포르를 포함한 13개 AWS 리전이 대상이며, 한 리전 안에서만 처리해야 하는 작업은 그 리전의 클러스터만 묶도록 안내합니다.

AP의 관점 · AP의 서비스 구현 교육에 참고한다면, 모의 서버 두 곳에서 지연·실패·전환·최종 오류를 나눠 기록하는 실험을 제안합니다. 진입점이 하나여도 모든 요청의 성공을 보장하지 않으며, 데이터 처리 지역과 재시도 경계를 별도로 확인할 수 있습니다.

공식 원문 읽기 ↗



AI 제작 · 편집 일러스트 · 실제 제품 화면 아님

03 · AI NEWS

HyperPod Slurm 확장, 설정 선언과 실제 적용 상태를 함께

AWS Blog · 발표 2026-11-20

AWS의 관리형 확장은 사용자·관측 설정을 새 노드에도 적용하고 노드별 건강 상태를 보고합니다.

AWS는 11월 20일 HyperPod Slurm의 관리형 클러스터 확장 사용법을 설명했습니다. 11월부터 제공한 관측·사용자 관리 확장은 API·콘솔에서 설정을 선언하면 스키마를 먼저 검증하고 새 노드·교체 노드에도 같은 버전과 설정을 적용합니다. 사용자·그룹별 SSH·sudo 권한을 노드 역할에 맞게 정할 수 있습니다. 노드별 충돌은 건강 상태에 나타나며, 확장 버전 업그레이드는 이전 버전으로 되돌리는 방식이 아니어서 유지보수 시간을 잡도록 안내합니다. 확장 자체의 추가 요금은 없지만 연결된 지표 저장소·비밀 관리 등 AWS 자원 비용은 별도입니다.

AP의 관점 · AP의 웹 서비스 운영 교육에 참고한다면, 가상 환경의 원하는 설정과 실제 적용 수·오류를 비교하는 점검표를 제안합니다. 설정 요청의 접수와 모든 대상에 적용된 상태를 구분하고, 계정별 최소 권한·변경 전 검증·복구 계획을 함께 살펴볼 수 있습니다.

공식 원문 읽기 ↗



AI 제작 · 편집 일러스트 · 실제 제품 화면 아님

04 · AI NEWS

Gemini의 대화 이력 가져오기, 새 발표보다 이전 절차 안내

Google Blog · 발표 2026-11-19

Google은 다른 AI 도구의 선호 요약과 대화 내보내기 파일을 Gemini로 가져오는 방법을 다시 안내했습니다.

Google의 11월 19일 글은 다른 AI 비서에서 맥락·선호·대화 이력을 가져오는 절차 안내입니다. Gemini 가져오기 설정에서 제시한 문구를 이전 도구에 넣어 선호 요약을 만든 뒤 그 결과를 Gemini에 붙여 넣는 방법을 설명합니다. 대화 이력은 이전 제공자의 계정 설정에서 ZIP으로 내보내고 Gemini 설정에 올리는 흐름입니다. 제목에 ICYMI가 붙은 재안내로, 모든 계정·국가에서 그날 새로 기능을 출시했다는 의미는 아닙니다. 실제 제공자의 내보내기 형식과 계정 설정을 확인해야 합니다.

AP의 관점 · AP의 AI 교육에 참고한다면, 실명·고객 자료를 담지 않은 가상 선호 요약으로 이전 전후 맥락을 비교하는 과제를 제안합니다. 대화 전체를 옮기는 일과 필요한 선호만 정리하는 일을 구분하고, 옮길 범위를 직접 검토하는 기준을 설명할 수 있습니다.

공식 원문 읽기 ↗



AI 제작 · 편집 일러스트 · 실제 제품 화면 아님

05 · AI NEWS

소규모 기업의 Workspace 활용, 고객 사례와 내 업무를 나누기

Google Blog · 발표 2026-11-19

Google은 문의 답변·교육 자료·생산 계획·마케팅 초안에 Gemini를 쓰는 네 기업 사례를 소개했습니다.

Google은 11월 19일 미국의 소규모 기업 네 곳이 Workspace와 Gemini를 활용한 사례를 공개했습니다. 행사 장비 대여 업체는 Gmail에서 할 일과 문의를 정리하고, 주조 업체는 기술 매뉴얼을 교육 자료로 바꾸거나 문서 초안을 만듭니다. 요구르트 업체는 Sheets의 생산 모델과 이메일 분류를, 음료 업체는 마케팅 구상과 자료 탐색을 활용한다고 설명합니다. 소개된 시간·매출 개선은 해당 고객의 보고이며 일반적인 효과를 입증하는 대조 시험은 아닙니다. 글은 별도의 AI 학습 자료와 무료 실시간 워크숍도 안내합니다.

AP의 관점 · AP의 콘텐츠 제작이나 Lia Select 운영에 참고한다면, 문의 초안·거래 조건·최종 답변을 나눈 가상 업무 실험을 제안합니다. 고객 사례의 수치를 목표 성과로 보장하지 않고, 우리 과제의 수정 횟수·누락 조건·소요 시간을 같은 기준으로 비교할 수 있습니다.

공식 원문 읽기 ↗

ONE THING TO TAKE AWAY

AP의 시선

불필요한 정보를 덜 싣는 것과 필요한 근거를 빼는 것은 다릅니다. 도구의 상세 정의는 필요할 때 읽더라도 권한·요청 결과·실제 적용 상태는 확인할 수 있어야 합니다. AP의 교육·콘텐츠·웹 서비스와 Lia Select 업무에 참고한다면, 가상 과제에서 선택한 도구·처리 위치·변경 결과·옮긴 정보의 범위를 함께 기록하는 실험을 제안합니다. 새 기능과 기존 절차 안내, 고객의 성과 사례를 나눠 읽으면 우리 업무에서 무엇을 따로 검증해야 할지 더 분명해집니다.