A DAILY PERSPECTIVE ON AI

AP *News*

생성 속도에서 검증 가능한 결과로

영상 생성·로컬 추론·문서 파싱의 효율과 다중 에이전트·과학 연구의 검증 조건을 살펴봅니다. 빠른 처리나 많은 시도를 곧바로 완성도와 협력의 효과로 읽지 않고, 실제 측정한 범위를 확인합니다.



AI 제작 · 편집 일러스트 · 실제 제품 화면 아님

THE DAILY BRIEF · 5 STORIES

전날인 11월 28일 이후 발표만으로 5건을 확보하지 못해 최근 7일로 범위를 넓혔습니다. Hugging Face에 실린 공식·팀 원문으로 11월 27일 1건, 26일 3건, 25일 1건을 선정했습니다. Hugging Face·ggml-org·NVIDIA·Ai2의 기술 사례와 연구를 구분했고 이전 호 원문·같은 이슈를 제외했습니다. AI 작성 및 원문 확인.



AI 제작 · 편집 일러스트 · 실제 제품 화면 아님

01 · AI NEWS

Trainium3 영상 생성 사례, 속도보다 조건과 연속성까지

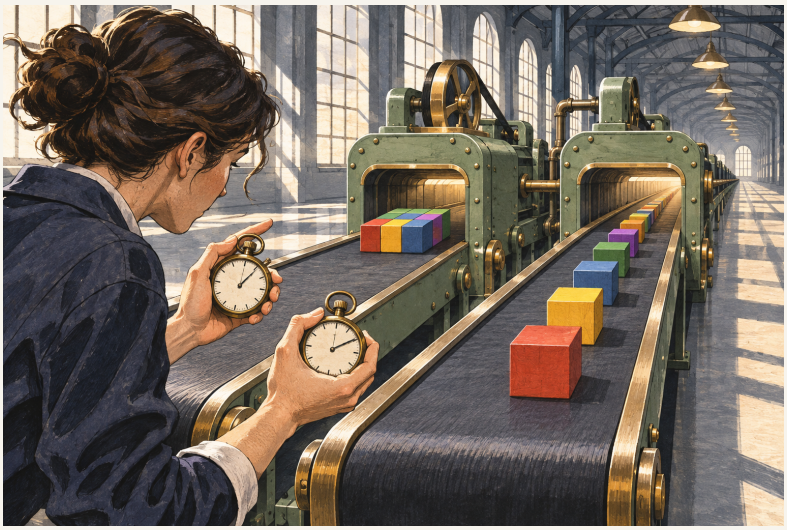
Hugging Face · 발표 2026-11-27

Hugging Face는 LTX-2.5의 영상·음향 생성 최적화와 장면 연결 과정을 공개했습니다.

Hugging Face는 11월 27일 AWS와의 협력으로 사용한 Trainium3에서 LTX-2.5를 최적화한 사례를 공개했습니다. 모델을 올린 뒤에는 장비 절반인 32 NeuronCores로 음향이 있는 10초·2048×1024 영상을 22~23초에 만들었다고 보고합니다. 모델 로딩은 약 3분, 첫 영상은 약 80초라는 조건도 구분합니다. 장면을 앞뒤 프레임으로 연결하고 여러 시도 중 결과를 골랐으며, 세부 오류·장면 사이 음향 불연속 때문에 초안 단계라고 설명합니다. 모든 장비·영상에 같은 속도나 완성도를 보장하는 비교는 아닙니다.

AP의 관점 · AP의 AI 콘텐츠 교육에 참고한다면, 가상의 짧은 이야기에서 핵심 장면을 먼저 정하고 인물·동작·소리의 연속성을 확인하는 과제를 제안합니다. 한 장면의 처리 시간과 작품 전체의 편집·검증 시간을 따로 기록할 수 있습니다.

공식 원문 읽기 ↗



AI 제작 · 편집 일러스트 · 실제 제품 화면 아님

02 · AI NEWS

llama.cpp의 AMD 대시보드, 입력 처리와 생성 속도를 나눈다

ggml-org · Hugging Face · 발표 2026-11-26

ggml-org는 빌드·모델·양자화별 로컬 추론 성능을 살펴보는 공개 대시보드를 소개했습니다.

ggml-org는 11월 26일 Hugging Face에서 AMD Ryzen·Radeon용 llama.cpp 벤치마크 대시보드를 공개했습니다. 입력 512토큰을 처리하는 PP512와 출력 128토큰을 만드는 TG128을 분리하며 두 기본 시험은 기존 문맥 깊이 0에서 시작합니다. 빌드 이력·양자화·문맥 증가에 따른 생성 처리량을 비교하고 원자료와 실행 흐름을 공개합니다. 당시 하드웨어 선택지는 Radeon 8060S와 Radeon AI Pro R9700S입니다. 토큰 처리량은 특정 기준 부하의 결과이며 모든 실제 대화·모델·장비의 응답 시간을 대신하지 않습니다.

AP의 관점 · AP의 AI·웹 서비스 교육에 참고한다면, 가상 성능 표에서 입력 길이·출력 길이·문맥·양자화 조건을 맞추는 과제를 제안합니다. 빠르다는 숫자 하나보다 어떤 작업을 얼마나 반복해 측정했는지 확인하는 기준을 세울 수 있습니다.

공식 원문 읽기 ↗



AI 제작 · 편집 일러스트 · 실제 제품 화면 아님

03 · AI NEWS

Nemotron 3.5 Parse, 문서 구조와 완료 여부를 함께 검증

NVIDIA · Hugging Face · 발표 2026-11-26

NVIDIA는 합성 자료·미세조정 방법·추론 가속 도구로 문서 파싱을 보완했습니다.

NVIDIA는 11월 26일 기존 Nemotron 3.5 Parse를 확장하는 자료와 방법을 공개했습니다. 문서 이미지를 읽기 순서·내용 종류·위치 상자와 표·차트의 구조화 기록으로 바꾸는 모델입니다. 합성 차트·표·다국어 OCR 자료 세 가지와 SFT 뒤 검증 가능한 보상으로 강화학습하는 방법, DFlash2 추론 도구를 설명합니다. 어려운 영역을 잘라 다시 읽는 다중 패스는 추가 처리량을 요구합니다. 학습 중간 체크포인트 비교·특정 GPU의 처리량·합성 양식 실험은 실제 고객 문서 전체의 정확도를 보장하는 수치와 구분해야 합니다.

AP의 관점 · AP의 뉴스레터·문서 교육에 참고한다면, 가상 표와 차트에서 원문 위치·행과 열·단위·완료 여부를 함께 확인하는 실습을 제안합니다. 빠른 변환 뒤 작은 영역을 다시 검증할 때 늘어나는 시간도 별도로 기록할 수 있습니다.

공식 원문 읽기 ↗



AI 제작 · 편집 일러스트 · 실제 제품 화면 아님

04 · AI NEWS

다중 에이전트 안내, 협력 효과와 더 쓴 자원을 구분

Hugging Face · 발표 2026-11-26

Hugging Face는 작은 공유 모델로 통신을 학습하는 실험과 공정한 비교의 조건을 설명했습니다.

Hugging Face는 11월 26일 다중 에이전트의 구축·학습·평가를 다룬 연재 첫 글을 공개했습니다. 예제는 Qwen3-0.6B 정책을 공유하는 네 에이전트가 각자만 아는 숫자를 통신해 정렬하는 과제입니다. 한 에이전트라도 전체 정답을 제출하면 팀 성공으로 판단하므로 모두의 정답을 요구하는 시험과 다릅니다. 이 과제는 필요한 정보를 교환하는 학습을 살펴보며, 같은 토큰 예산의 단일 에이전트보다 협력이 낫다는 직접 증명은 아니라고 명시합니다. 더 많은 시도와 통신 자체의 비용도 분리해야 합니다.

AP의 관점 · AP의 AI 교육·서비스 설계에 참고한다면, 실제 유료 실행 없이 가상 역할별 자료와 메시지를 기록하는 과제를 제안합니다. 결과의 정확성뿐 아니라 누가 어떤 정보를 받았는지, 추가 통신과 재시도가 얼마나 늘었는지 함께 비교할 수 있습니다.

공식 원문 읽기 ↗



AI 제작 · 편집 일러스트 · 실제 제품 화면 아님

05 · AI NEWS

과학자 17명 인터뷰, AI 위임 뒤에도 판단 책임은 남는다

Ai2 · Hugging Face · 발표 2026-11-25

Ai2는 연구 에이전트의 검증·개인정보·객관성·책임·투명성에 관한 질적 조사 결과를 소개했습니다.

Ai2는 11월 25일 자연·응용·사회과학 분야 과학자 17명과의 인터뷰를 소개했습니다. 연구진은 과학적 신뢰의 다섯 차원으로 검증·개인정보·객관성·책임·투명성을 정리했습니다. 문헌 요약의 출처뿐 아니라 빠뜨린 반대 근거·사용한 방법·접근한 자료도 확인해야 한다는 우려를 다룹니다. 과학자가 평가 기준을 함께 정하는 방향을 제안하지만 아직 개발·시험할 구상이라고 구분합니다. 이 질적 인터뷰가 모든 연구자의 의견이나 AI 도입 효과를 대표하는 대조 실험은 아닙니다.

AP의 관점 · AP의 AI 교육·콘텐츠·Lia Select 조사에 참고한다면, 가상 조사 결과에서 결론을 지지하는 자료와 반대 자료·누락을 함께 확인하는 과제를 제안합니다. 답이 그럴듯한지에 더해 자료 권한·방법 재현·최종 판단 책임을 기록할 수 있습니다.

공식 원문 읽기 ↗

ONE THING TO TAKE AWAY

AP의 시선

성능 숫자와 자동화 사례는 장비·작업·비교군·성공 정의가 함께 있어야 판단 근거가 됩니다. AP의 AI 교육·뉴스레터·웹 서비스와 Lia Select 조사에 참고한다면, 가상 적용안마다 처리 시간·재시도·검증 비용·누락 근거·판단 주체를 나눠 적는 실습을 제안합니다. 한 번의 정답이나 빠른 시연을 실제 운영의 완성으로 확대하지 않고, 확인된 결과에서 다음 단계를 정할 수 있습니다.