Weekly AI NEWS

PseudoLab

SOTA AI Review Project

AI NEWS

2025.09.29 - 2025.10.05

퍼플렉시티, AI 웹 브라우저 '코멧' 전 세계 무료 공개

10위

- 퍼플렉시티 AI 웹 브라우저, **코멧 전 세계 무료 공개**
- **코멧 기능**
  - **(무료) 사이드카 어시스턴트**: 브라우저 사이드의 AI 챗봇 -> 웹 페이지 요약, 질의응답, 일부 에이전트 기능 (ex. 콘텐츠 관리와 탐색 등)
    + 디스커버 (맞춤형 뉴스 추천), 스페이스 (프로젝트 관리), 쇼핑, 트래블, 파이낸스, 스포츠 등
  - **(맥스 구독자) 이메일 어시스턴트**: 받은 이메일함 정리 우선순위화, 사용자톤의 회신 작성, 일정 관리 등
    + 고성능 AI 모델 활용, **퍼플렉시티 신제품 및 신 기능 우선 사용**

    (ex. 백그라운드 어시스턴트: 대시보드에서의 멀티 태스크 동시 자동 관리 어시스턴트)

출처: https://www.aitimes.com/news/articleView.html?idxno=202946

9위

메타, 자체 칩 제작 위해 스타트업 리보스 인수 계획

Meta × Rivos

- 메타, 인공지능(AI) 인프라 핵심인 반도체 제작 역량 강화를 위해 **미국 반도체 스타트업 리보스(Rivos) 인수 추진**
- **리보스**: **RISC-V 오픈 소스 아키텍처**를 기반으로 **GPU 및 AI 가속기를 설계하는** 칩 스타트업
  → SoC 및 PCIe 가속기 등 칩 설계 역량 보유
  (* 8월 기준 약 20억 달러 가치 평가)
- **인수 배경**
  ① **NVIDIA GPU 조달에 매년 수십억 달러 지출**
  → **AI 모델 학습 및 추론 비용 절감**을 위해 **맞춤형 반도체 개발**을 최우선 과제로 삼음.
  ② **자체 AI 칩 프로젝트 MTIA를 진행**(-> 칩 생산 및 데이터센터 시범 적용) **중**이나, 마크 저커버그 CEO는 **개발 속도가 기대에 못 미친다** 판단하여 외부 전문 인력 영입 필요성 검토
- **목표**: 리보스 인수를 통해 **맞춤형 실리콘 개발 노력을 가속화**할 계획

출처: https://www.aitimes.com/news/articleView.html?idxno=202903

번외

"아마존·구글, 젠슨 황에 자체 AI 칩 발표 내용 미리 보고"

- 아마존과 구글이 자체 AI 칩 개발 계획을 발표하기 전, **NVIDIA 젠슨 황 CEO에게 먼저 정보를 보고**하는 것으로 알려짐
- **배경 및 의미**
  ① NVIDIA GPU 없이는 아마존과 구글의 클라우드 사업이 존재할 수 없을 정도로 **AI 산업 내 엔비디아의 막대한 영향력**을 보여줌.
  ② 자체 칩 개발에 대해 황 CEO는 **공개적으로 관용적**이지만, **기업들이 예상치 못한 행보를 보이는 것을 원치 않기 때문**에 "남다른 경의를 표현"하는 것임.
- **NVIDIA의 영향력 확대**
  - 엔비디아는 주요 클라우드 업체 및 AI, 클라우드, 로봇 스타트업 (오픈AI, 코어위브, 미스트랄 AI 등)에 투자 → 업계 영향력 확대
  - 막대한 매출로 인해 시가 총액이 4조 5000억 달러를 돌파하는 등 막강한 재정적 지위 확보

출처: https://www.aitimes.com/news/articleView.html?idxno=202908

8위

앤트로픽, 30시간 코딩 가능한 '클로드 4.5' 출시...기업 시장 조준

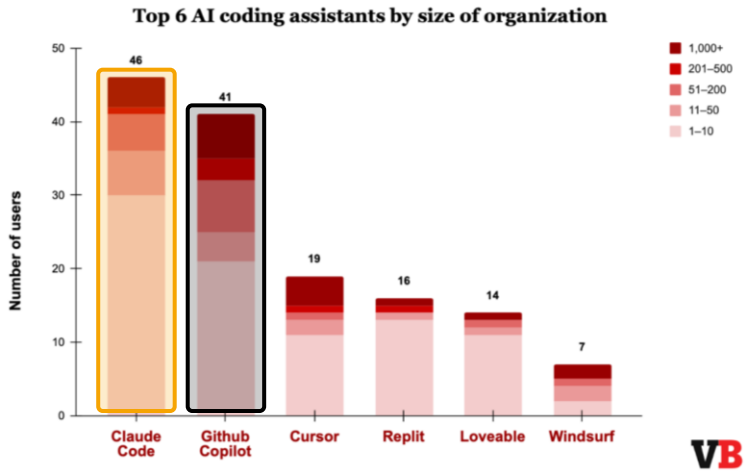
	Claude Sonnet 4.5	Claude Opus 4.1	Claude Sonnet 4	GPT-5	Gemini 2.5 Pro
Agentic coding *SWE-bench Verified*	**77.2%** **82.0%** with parallel test-time compute	74.5% 79.4% with parallel test-time compute	72.7% 80.2% with parallel test-time compute	72.8% GPT-5 74.5% GPT-5-Codex	67.2%
Agentic terminal coding *Terminal-Bench*	**50.0%**	46.5%	36.4%	43.8%	25.3%
Agentic tool use *τ2-bench*	Retail **86.2%** Airline **70.0%** Telecom **98.0%**	Retail **86.8%** Airline 63.0% Telecom 71.5%	Retail 83.8% Airline 63.0% Telecom 49.6%	Retail 81.1% Airline 62.6% Telecom 96.7%	— — —
Computer use *OSWorld*	**61.4%**	44.4%	42.2%	—	—
High school math competition *AIME 2025*	100% (python) 87.0% (no tools)	78.0%	70.5%	99.6% (python) 94.6% (no tools)	88.0%
Graduate-level reasoning *GPQA Diamond*	83.4%	81.0%	76.1%	85.7%	**86.4%**
Multilingual Q&A *MMMLU*	89.1%	**89.5%**	86.5%	89.4%	—
Visual reasoning *MMMU (validation)*	77.8%	77.1%	74.4%	**84.2%**	82.0%
Financial analysis *Finance Agent*	**55.3%**	50.9%	44.5%	46.9%	29.4%

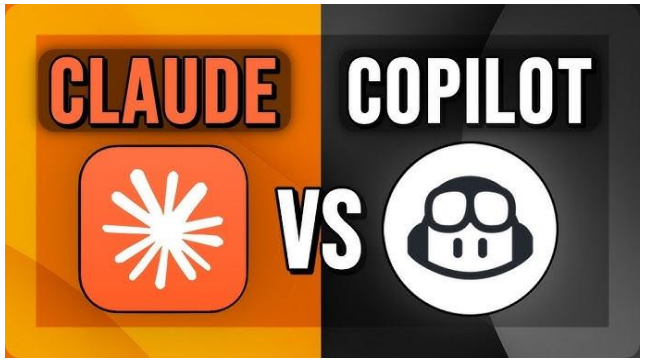
출처: https://www.aitimes.com/news/articleView.html?idxno=202843

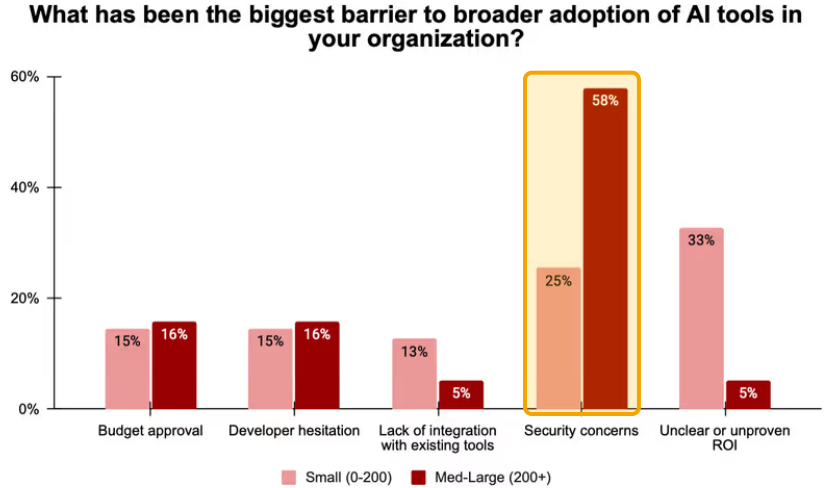
- 앤트로픽, 기업용 AI 시장 점유율 강화 목표로 **'클로드 소네트 4.5(Claude Sonnet 4.5)'** 공개
- **핵심 성능 및 기능**
  - **코딩 지속 시간**: 코딩 작업 지속 시간 최대 30시간으로 4배 이상 증가
    → 복잡한 기업 프로젝트 자동화에 유리
  - **용도 및 기능**: 에이전트 기능 개선 (→ 코드 실행, 문서 생성 등 챗봇 앱에서 직접 처리 가능)
    + VS 코드 연동, 장기 작업 메모리 등 추가
    ⇒ 사이버보안, 금융 등 B2B 영역 적합
  - **성능**: 'SWE-벤치 베리파이드'에서 77.2%를 기록
    → 소프트웨어 엔지니어링 분야 최첨단(SOTA) 성능 입증
- **안전 및 시장 전략**
  - **AI 안전**: 자체 'ASL-3' 등급 적용 + 안전 조치 강화
    (ex. 화학/핵무기 관련 위험 입력 감지 필터 및 모델의 기만 행위 대응 등)
  - **개발자 지원**: '클로드 에이전트 SDK' 및 연구용 미리보기, 'Imagine with Claude' 제공
  - **가격**: 성능 향상 but 가격 동결 = API 사용료는 소네트 4와 동일
    (→ 입력 100만 토큰당 $3, 출력 100만 토큰당 $15)

'클로드 코드'의 대기업 채택률이 '깃허브 코파일럿'보다 낮은 이유는

7위





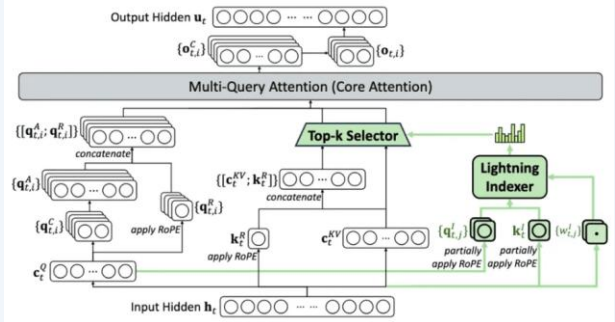


- **코딩 AI 툴 시장 점유율**: Claude Code (46%) vs. GItHub Copilot (41%) ...
- **코딩 AI 툴 성능 평가**
  - 최초 작성 시간 (Time-to-First Code) 평가: 17초(Github Copilot) <<< 36초 (Claude Code) ---> 원인: Claude의 오류 방지 기능 (* 정확도 – 출력 시간 trade-off)
  - 6가지 항목 평가 (보안, 배포 유연성, 통합 기능, 코딩 성능, 비용 예측 가능성, 기업 지원
    : Github Copilot (보안 - 중간, 배포 유연성 – 낮음, **나머지 4개 분야 높음**) <<< CLAUDE Code **(보안 – 높음,** 나머지 5개 분야 - 중간)
    ↪ 가장 기업 친화적인 코딩 도구로 평가!
- **기업 (대기업 / 스타트업) 고려 사항** : 보안, 규정 준수, 배포 제어 -> ① 대기업: 규정 준수 실패 및 사고 ⇒ GitHub Copilot 선호 (82%)!
  ② 소규모 기업: 가격의 합리성 ⇒ Claude Code, Cursor 등 **최신 도구 선호!**
- **코딩 AI 툴 시장 현황 및 전망**
  - 현황: 멀티 툴: 두 개 이상의 코딩 도구 사용 중 (49%) -> GitHub Copilot + Claude Code 사용 중 (26%)
  - 평가: (Claude Code) 멀티 모델 사용 X -> 기업 요구사항 불만족
    (전체) 어떤 코딩 AI 툴도 기업의 포괄적 요구 모두 만족 X -> 기업 요구를 수용하는 플랫폼의 성장 가능성 ↑

출처: https://www.aitimes.com/news/articleView.html?idxno=202957
https://venturebeat.com/ai/github-leads-the-enterprise-claude-leads-the-pack-cursors-speed-cant-close

6위

딥시크, API 비용 절반으로 줄인 실험 모델 'V3.2' 출시
딥시크의 새로운 'V3.2' 모델은 중국산 칩 지원용으로 개발



Benchmark		DeepSeek-V3.1-Terminus	DeepSeek-V3.2-Exp
General	MMLU-Pro	85.0	85.0
	GPQA-Diamond	80.7	79.9
	Humanity's Last Exam	21.7	19.8
Search Agent	BrowseComp	38.5	40.1
	BrowseComp-zh	45.0	47.9
	SimpleQA	96.8	97.1
Code	LiveCodeBench	74.9	74.1
	Codeforces-Div1	2046	2121
	Aider-Polyglot	76.1	74.5
Code Agent	SWE Verified	68.4	67.8
	SWE-bench Multilingual	57.8	57.9
	Terminal-bench	36.7	37.7
Math	AIME 2025	88.4	89.3
	HMMT 2025	86.1	83.6

딥시크, 연산 효율 개선 및 비용 절감 위한 새로운 실험적 LLM(-> 차세대 아키텍처를 향한 중간 단계), "딥시크-V3.2-Exp" 출시

1. 기술 및 목적

- 목적: 희소 어텐션(Sparse Attention) 구조 도입 → 장문 문맥 처리 시 추론 비용 절반 이상 절감하는 것
- 원리: '라이트닝 인덱서'와 '세밀 토큰 선택 시스템' 활용 문맥에서 중요 토큰만 선 → 제한된 희소 어텐션 창에 집중
  ⇒ 연산 부담 크게 감소
- 학습 기법: 분야별 특화 모델을 '전문가 증류(specialist distillation)' 기법으로 통합
  → 이후 GRPO 기반의 단일 강화 학습(RL) 방식 도입
  ⇒ 추론, 에이전트 성능, 휴먼 얼라인먼트 성능 향상

2. 성능 및 가격 경쟁력

- 성능: 이전 버전과 전반적 동일 or 소폭 개선
  (→ 'MMLU-Pro' 85.0 유지, 'AIME 2025' 89.3 상승, 코딩 관련 '코드포스' 2121로 개선)
- 가격: API 호출 비용 최대 50%까지 할인 (-> 경쟁사 대비 비싼 가격)
  → 입력 캐시 미스 비용 0.56달러 → 0.28달러, 출력 비용 1.68달러 → 0.42달러

3. 중국 AI 칩과의 관계

- 배경: 딥시크V3.1의 화웨이 어센드 칩 및 CANN SW 스택 최적화 + 화웨이의 추론 레시피 및 캠브리콘의 호환 기술
  ⇒ 엔비디아 CUDA 스택에서 벗어난 첫 전용 모델
- 목적: 중국 칩의 엔비디아 칩 대비 낮은 성능 -> 모델의 성능 향상보다 효율성 집중

출처: https://www.aitimes.com/news/articleView.html?idxno=202864
https://www.aitimes.com/news/articleView.html?idxno=202896

상하이교통대, 78개 사례 학습으로 AI 자율성 대폭 향상..."데이터, 양보다는 질"

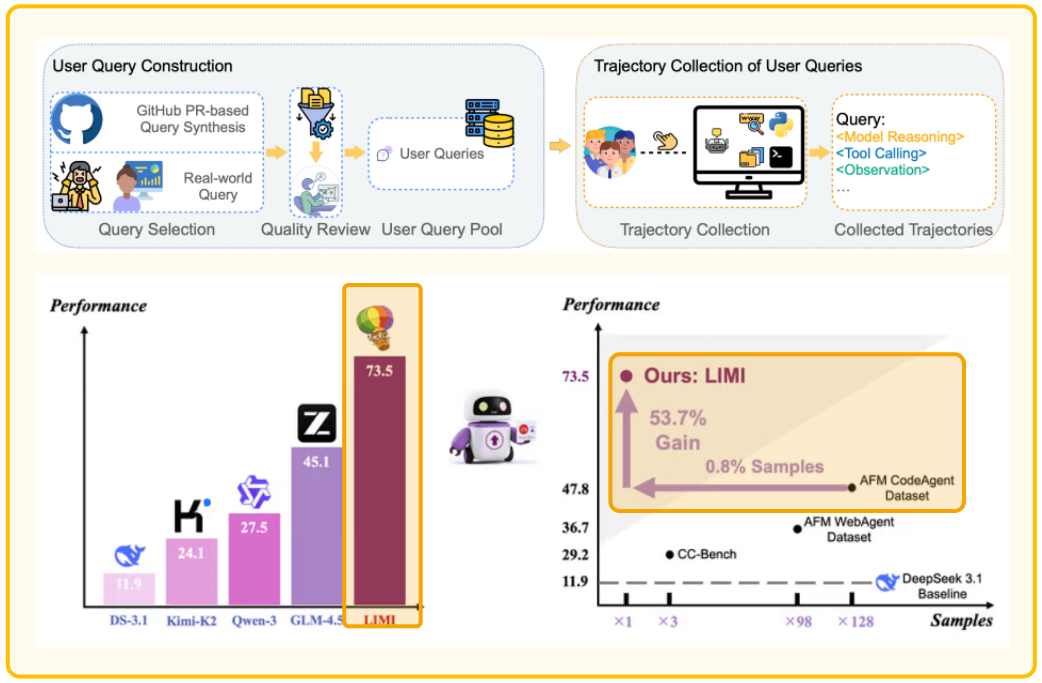
5위

1. 연구 개요

- **개요:** 상하이교통대학교 - 중국 생성 AI 연구소(GAIR) 공동 연구팀, 'LIMI(Less Is More for Intelligent Agency)' 공개
  ⇒ LLM의 자율적 행동 능력(AI 에이전트 기능) 확보는 선별된 고품질 시연(demonstration)이 핵심
- 목표: 소량의 데이터만으로도 고도화된 자율적 지능(agency)을 학습 가능한 데이터 구성 방식 구현

2. 실험 결과 및 성능 향상

- **데이터 효율성:** 단 78개의 고품질 예시 데이터만으로 수천 개의 사례로 훈련된 모델을 능가하는 성능 구현.
  + 1만 개 샘플로 훈련된 모델보다 128배 적은 데이터로도 53.7% 향상된 성능 구현 (코딩, 도구 사용, 과학 연산 등 전 영역에서 비교 모델 능가)
- **성능:** 에이전트 성능 벤치마크 **'에이전시벤치(AgencyBench)' 기준 LIMI 미세조정한 모델이** 평균 73.5% 기록 → 기존 최고 성능 모델(GLM-4.5, 45.1%)을 크게 앞섬



3. LIMI 프레임워크의 데이터 구성 혁신

- **학습 시연 구성:** '질의(Query)' + '과정(Trajectory)'
  - **질의:** 사용자의 **요청**
  - **과정:** AI가 문제를 해결하기 위해 수행하는 **전체 문제 해결 과정**
    = 문제 해결 시도와 실패, 학습 개선 전략을 조정하는 전 과정의 기록
    (ex. 단계별 사고 전개, 코드 작성 및 실행, 외부 도구 호출, 오류 발생 시 수정 과정 등)
- **학습 시연 데이터 수집:** 실무 개발자와 연구자의 **실제 문제 60개 수집** 후
  → GPT-5를 이용해 **질의 합성**
  → **컴.공 박사과정 4명의 직접 검증**을 거쳐 최종 78개 샘플 확정

4. 시사점

- **개발 방식의 재정의:** 자율형 AI 개발 → **'얼마나 현명하게 시연을 구성하는가'**
- **기업 환경에의 의미:** 데이터 수집 비용이 높은 기업 환경
  → **소규모 고품질 데이터셋 구축** 통한 **맞춤형 AI 에이전트 효율적 개발**

출처: https://www.aitimes.com/news/articleView.html?idxno=202979

4위

알트먼, 2033년까지 컴퓨팅 용량 100배 규모인 250GW로 확대

출처: https://www.aitimes.com/news/articleView.html?idxno=202860

1. 목표 및 동기

- **동기:** 빅테크 간 컴퓨팅 경쟁 중 선두 유지 위한 인프라 확장
- **목표:** OpenAI, 2033년까지 250기가와트(GW)의 **컴퓨팅 용량 확보** 계획
- **의미:** 250GW = **원자력 발전소 약 250개**가 필요한 규모 (→ 12조 5,000억 달러)
  = 미국 전체 전력 수요의 3분의 1
  = 현재 칩 기준 **GPU 약 1억 장 필요** 예상
  = 현재 **예상 용량(2.4GW)의 100배** 이상 규모

2. 투자 및 인프라 구축 현황

- **목표:** (단기) 금년 2.4GW 증설 → 연말까지 100만 개 이상의 GPU 온라인 연결 예정
  (장기) **매주 1GW의 새로운 AI 인프라를 생산할 수 있는 공장을 만드는 것** 목표
- **투자:** 엔비디아 1,000억 달러 파트너십, 스타게이트 4,000억 달러 투자,
  코어위브와의 계약 확대 등
- **인프라 구축: 스타게이트 1** = 1.1GW 규모 데이터 센터
  (→ 엔비디아 최신 칩 40만 장 탑재 예정, 텍사스 건설 중)
  + 2028년까지 **8GW 규모의 인프라 구축** 엔비디아와 합의

3. 효과 및 전망

- **OpenAI의 경쟁 일시적 우위: 2028년까지 최소 8GW, 2033년까지 250GW**
  (→ 글로벌 클라우드 사업자 2위, MS 애저의 2023년 전체 용량(5GW) 추월)
- **업계 동향: 인프라 증설 경쟁 심화** → 경쟁사들의 8~10GW의 데이터센터 추가 계획 발표

3위

오픈AI, '소라 2'·소셜 앱 '소라' 출시..."동영상 챗GPT 순간"

'소라', 출시 이틀 만에 미국 iOS 3위 기록...'성공 조짐'

- OpenAI, 최신 비디오 생성 모델 '소라 2(Sora 2)' 출시, 이를 탑재한 새로운 iOS 소셜 앱 '소라'

- **1. 소라 2 모델**
  - 성능 향상: Sora 1 대비 **물리적으로 더 정확하고 사실적 + 제어 용이**
    → 주변 환경과의 물리적 이해가 필요한 복잡한 동작 표현 가능
  - 기능: ① 생성된 영상에 대사, 음향 효과 추가 가능
    ② 기존 영상에 새로운 객체를 추가 가능

- **2. 소셜 앱 '소라'**
  - **기능: 카메오(Cameo):** 사용자가 소라로 생성된 모든 영상에 자신(사용자)을 넣을 수 있는 기능
    (* 본인 확인 및 영상/음성 파일 업로드 필요)
    추천 알고리즘: 활동, IP 주소, 이전 게시물 참여 내역, 챗GPT 대화 내역 등 기반 추천 제공
  - **저작권**: 할리우드 주요 스튜디오들에게 자사 캐릭터가 생성 영상에 포함되지 않도록 하는 옵션 제공
  - **출시 계획: 미국 캐나다 우선 출시 / ChatGPT Pro 사용자**는 '소라 2 프로' 모델 우선 사용 가능
  - **가격 정책: 무료 서비스**(→ 수요가 많은 시기에 추가 영상 제작 과금 방안 검토 중)

- **3. Sora 앱 초반 반응**
  - **앱 다운로드 수:** 9/30 – 10/1 이틀 간 **총 16만4000건**의 설치
    → 미국 iOS 순위 3위 / 역대 AI 앱의 2일차 iOS 순위 2위
    (* 현재 Sora 앱은 초대 받은 사용자 및 그의 초대 코드 4개 수령자만 사용 가능)
  - **앱 평점: 긍정적** → iOS 평점 4.0/5.0

출처: https://www.aitimes.com/news/articleView.html?idxno=202885&page=2&total=8805
https://www.aitimes.com/news/articleView.html?idxno=202954

화웨이, LLM 경량화 기술 'SINQ' 오픈 공개..."메모리 사용 70%까지 절감"

2위

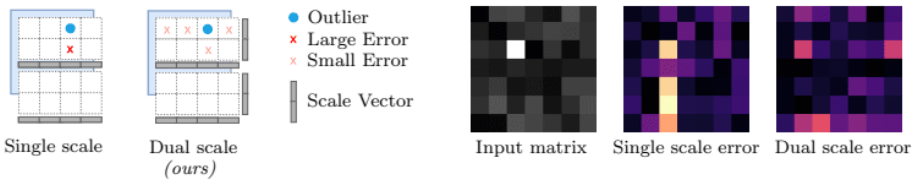
1. 기술 SINQ (Sinkhorn-Normalized Quantization) 개요

- 의미: 기존 양자화의 정밀도 문제를 극복하고 보정 데이터가 필요 없는 '플러그 앤 플레이(plug-and-play)' 솔루션
- **목표:** 대형언어모델(LLM)의 메모리 요구량 크게 줄이면서도 출력 품질 유지 ⇒ 고성능 GPU 없이도 모델의 효율적 축소 및 배포가 가능하도록 하드웨어 진입 장벽 낮추기

2. 성능 및 비용 효율성

- **메모리 절감:** 모델 크기와 비트폭에 따라 **메모리 사용량을 60~70%까지 절감 가능** (예: 60GB 필요 모델을 20GB 환경에서 구동).
- **하드웨어 비용 절감:** 엔비디아 A100/H100 급 GPU 없이도 **GeForce RTX 4090 한 장 등으로 LLM 실행 가능.**
- **클라우드 비용 절감:** A100 기반 작업 대비 24GB GPU **인스턴스 비용이 1/3 수준**으로, 장기 추론 작업에서 큰 비용 절약 효과.

	Method	Qwen3-14B					
		AIME 2024		*AIME 2025*		Avg.	
		Tok.	Acc. (%)↑	Tok.	Acc. (%)↑	Δ Tok.	Acc. (%)↑
	Original (FP16)	*11 464*	*76.70*	*12 636*	*63.30*	*0*	*70.00*
CALIBRATION-FREE 4-BIT	RTN	10 973	66.70	12 642	50.00	-242	58.35
	BnB (FP4)	11 500	60.00	12 455	53.30	-72	56.65
	BnB (NF4)	12 132	70.00	12 899	56.70	+930	63.35
	Hadamard + RTN	11 210	70.00	12 989	53.30	+99	61.65
	HQQ	11 862	70.00	12 991	56.70	+367	63.35
	SINQ	11 660	**73.30**	12 305	**63.30**	-67	**68.30**



3. SINQ의 핵심 혁신 기술

① 이중 축 스케일링 (Dual-Axis Scaling):

- 한 행렬의 **행(row)과 열(column)에 각각 별도의 스케일 벡터** 적용
- 이를 통해 이상치(outlier)의 영향 감소 + 행렬 내 오차 분포 정밀 조정
  → 낮은 비트폭에서도 높은 정밀도를 유지

② 싱크혼-크노프 정규화 (Sinkhorn-Knopp Normalization):

- **싱크혼 반복 알고리즘 기반**으로 **행과 열 단위로 표준편차를 빠르게 조정**한다.
- 행렬 불균형(matrix imbalance) 지표 최소화
  → 기존의 비보정(calibration-free) 양자화 기술 대비 효율적

4. 벤치마크 결과

- 정밀도: 기존 비보정 양자화 기술들(ex. RTN, HQQ 등)을 능가
  + 일부 구간에서는 보정 기법과 유사한 성능
- 런타임 효율: 양자화 수행 속도가 HQQ보다 2배, AWQ보다 30배 이상 빠름

출처: https://www.aitimes.com/news/articleView.html?idxno=202980

1위

캘리포니아주, 기업 반대에도 첨단 AI 규제 법안 통과

- 캘리포니아 주, 대형 AI 기업에 **안전성과 투명성**을 요구하는 법안 **'SB 53'** 최종 서명
- 규제 대상: **연 매출 5억 달러 이상의 주요 AI 기업**
  (ex. 오픈AI, 앤트로픽, 메타, 구글 딥마인드 등)
- 주요 의무 사항
  ① **재앙적 위험 평가 및 공개**: AI 모델이 인간 통제를 벗어나거나 생물무기 개발에 악용될 가능성 등 재앙적 위험에 대해 평가하고, 결과를 공개해야 함.
  ② **중대 안전사고 보고 체계**: 사이버 공격, 모델의 기만적 행위 등 중대한 안전사고 발생 시, 캘리포니아 비상 서비스국에 보고 가능한 공공 보고 체계 마련
  ③ **내부 고발자 보호**: 잠재적 위험을 제기한 내부 직원에 대한 보호 장치 보장 위반 시 처벌
  ⇒ 의무 위반 기업에 **최대 100만 달러의 벌금 부과** 가능
- 업계 반응
  - 긍정: 앤트로픽 – 잭 클라크 공동 창립자, "공공 안전과 혁신을 조화시킨 강력한 틀"
  - 부정: 오픈AI, 메타, 법안 반대 로비 진행 + 오픈AI, 주지사에게 서명 보류 요청
- 비판: **주 단위 규제가 파편화**되어 스타트업에 부담을 줄 수 있다는 우려 제기
- 향후 논의: 연방 차원에서도 양당 의원들이 **AI 규제 표준 마련의 필요성**에는 공감,
  그러나 **주 규제와의 관계 설정**에 이견이 있어 논의 중
  테드 리우 민주당 의원,
  "AI 규제는 피할 수 없다."
  "다만 17개 주가 따로 할 것인지, 의회가 할 것인지 선택해야 한다."

출처: https://www.aitimes.com/news/articleView.html?idxno=202865