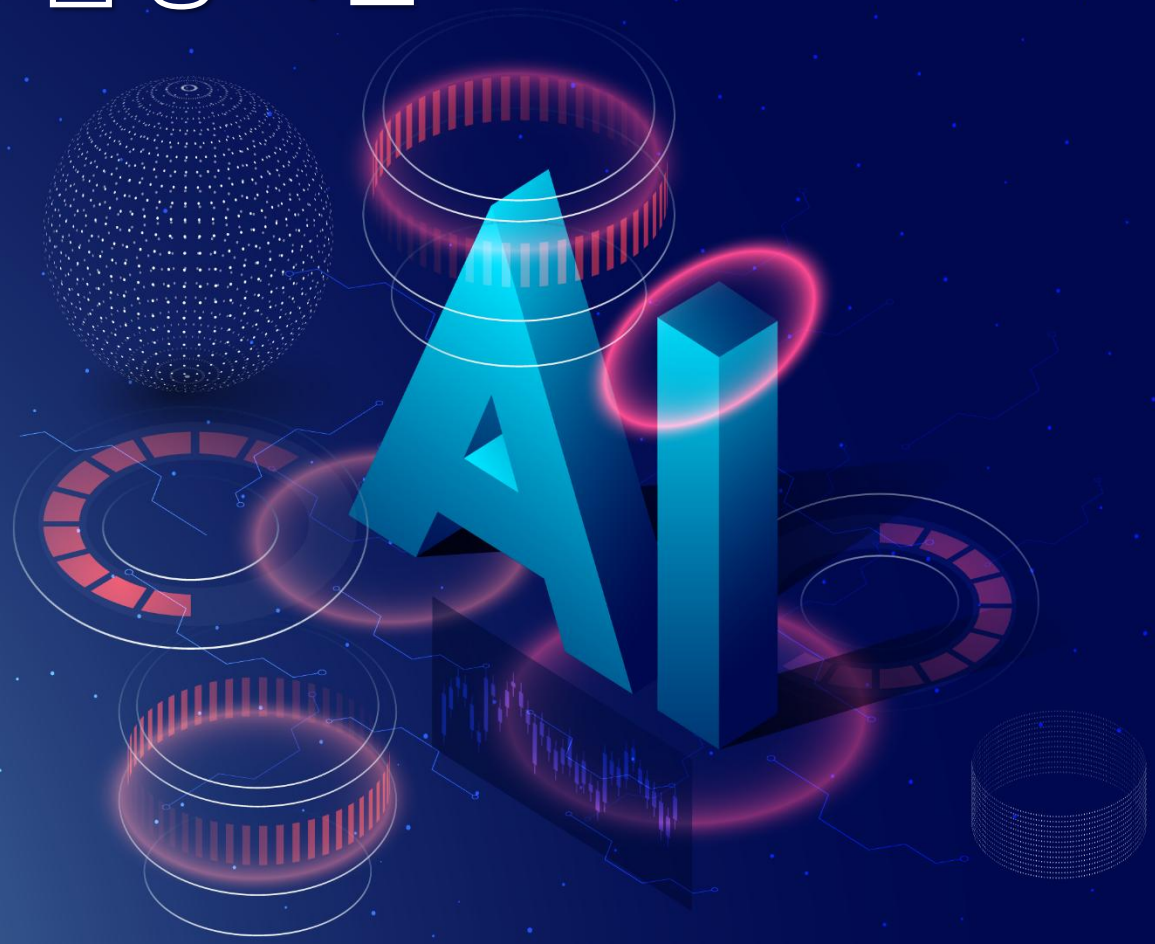
Research Center | 2026.04.07

Daishin 증권

AI Monthly

Agentic AI 사용에 급증하는 CPU 수요



▪ 서영재 youngjae.seo@daishin.com

Contents

I. AI 산업 트렌드 05

II. AI 뉴스 & 데이터 대시보드 18

Summary

Agentic AI 사용에 급증하는 CPU 수요, 단기 이슈에 그칠 Turboquant

- Agentic AI 시대가 도래하면서 CPU 수요가 증가하기 시작
  단순히 질문에 답을 하는 '채팅'의 단계를 넘어, AI가 스스로 목표를 세우고 도구를 사용하며 결과를 도출하는 '실행'의 단계
  이 과정에서 전체 시스템을 조율하고 복잡한 명령을 내리는 CPU의 가치는 오히려 더 커지게 됨
  CPU 수요가 증가하고 있다는 신호는 AMD, 인텔, Arm 직전 분기 실적 발표부터 나타남

- ARM 경영진은 Arm Everywhere 행사를 통해 장기 전략을 제시. 자체 설계한 칩은 Arm AGI CPU는 2026년 하반기부터 출하할 예정
  Arm 아키텍처 기반 CPU는 AMD, 인텔의 x86 아키텍처 기반 CPU 대비 저전력이라는 특성을 바탕으로 하이퍼스케일러 기준 점유율 50% 상회
  1GW AI 데이터센터 건설을 기준으로 Arm CPU를 사용할 경우 x86 CPU 대비 100억 달러를 절감할 수 있다고 주장
  Arm은 이를 바탕으로 2031년 칩 판매 매출이 150억 달러를 상회할 것으로 전망

- 구글 리서치가 대규모 언어 모델(LLM)의 KV 캐시와 벡터 검색 엔진의 메모리 병목 현상을 해결하기 위해 TurboQuant, QJL, PolarQuant 도입
  해당 기술들은 기존 방식 대비 최대 8배의 연산 속도 향상과 6배 이상의 메모리 절감 효과를 입증
  TurboQuant 발표 이후 메모리 기업들 주가 급락. 메모리를 효율적으로 사용할 경우 수요가 줄어들 수 있다는 우려 때문
  그러나 1) TurboQuant 구현 기술이 이미 1년 전에 발표됐다는 점, 2) 제번스의 역설로 효율적 사용이 실제 수요를 더 늘릴 것이라는 반박 등이 제기
  명확한 것은 메모리를 효율적으로 사용하려는 시도가 이루어짐에도 Agentic AI 등장으로 메모리 수요가 지속적으로 증가하고 있다는 것

- 지난해 말 발표된 Claude Opus 4.5를 기점으로 Agentic AI 시대 도래
  OpenRouter의 주간 토큰 생성량을 보면 지난해 말부터 급격하게 증가. 토큰 생성량은 지난해 말 대비 4배 가까이 증가한 것으로 추정
  토큰 생성량 증가를 보면 클라우드 기업들의 CapEx는 더욱 증가할 가능성이 높다고 판단
  주요 클라우드 기업들의 실적 발표가 2월 초에 이루어진 점을 고려하면 CapEx 전망치에 지금의 토큰 생성량이 반영되지 않았을 가능성 존재

- 한편, LLM Leaderboard를 보면 중국의 LLM들이 상위권을 차지
  쉽고, 간단한 태스크는 API 가격이 낮은 중국 모델을 쓰는 결과
  어렵고 복잡한 태스크의 경우에는 클로드, 제미나이, 그록을 쓰는 모습을 보임

I. 이달의 AI 산업 트렌드: 메모리 효율화 & 새로운 병목 CPU

① 구글 리서치의 TurboQuant

② 새로운 병목으로 떠오르는 CPU

③ Agentic AI 수요 증가 가속화

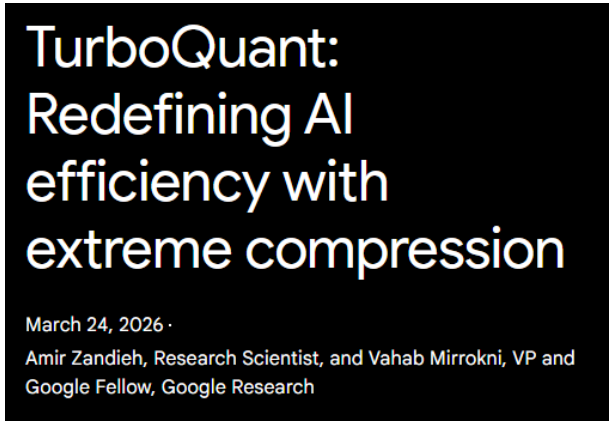
④ 온디바이스 AI 신호탄이 될 Gemma 4

1. AI 산업 트렌드

① 구글 리서치는 TurboQuant, PolarQuant, QJL 도입 → 기존 대비 8배 연산 속도↑, 6배 메모리 절감 달성

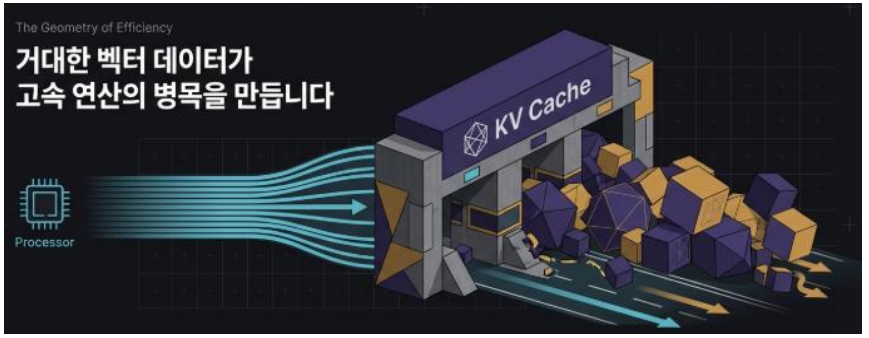
- 구글 리서치는 대규모 언어 모델(LLM)의 KV 캐시와 벡터 검색 엔진의 메모리 병목 현상을 해결하기 위해 TurboQuant, QJL, PolarQuant 도입
  해당 기술들은 기존 방식 대비 최대 8배의 연산 속도 향상과 6배 이상의 메모리 절감 효과를 입증
- 벡터는 AI 모델이 정보를 이해하고 처리하는 근본적인 방식. "고차원" 벡터는 이미지의 특징, 단어의 의미 또는 데이터셋의 속성과 같은 복잡한 정보 포착
  고차원 벡터는 매우 강력하지만 방대한 양의 메모리를 소비하며, 이는 키-값(KV) 캐시의 병목 현상으로 이어짐
  KV 캐시는 컴퓨터가 느리고 거대한 데이터베이스를 뒤지는 대신 즉시 검색할 수 있도록 자주 사용하는 정보를 간단하게 저장하는 고속 "디지털 치트 시트"
- 벡터 양자화는 고차원 벡터의 크기를 줄이는 강력하고 고전적인 데이터 압축 기술. 이 최적화는 AI의 두 가지 핵심 측면을 해결
  첫째, 대규모 AI 및 검색 엔진을 구동하는 고속 기술인 벡터 검색에서 유사도 조회를 더 빠르게 하여 검색 성능을 강화
  둘째, 키-값 쌍의 크기를 줄여 KV 캐시 병목 현상을 해소함으로써 더 빠른 유사도 검색을 가능하게 하고 메모리 비용을 낮춤
- 그러나 전통적인 벡터 양자화는 대개 자체적인 "메모리 오버헤드"를 발생시킴. 데이터의 작은 블록마다 양자화 상수를 계산하고 저장해야 하기 때문
  TurboQuant는 벡터 양자화의 메모리 오버헤드 문제를 최적으로 해결하는 압축 알고리즘. ICLR 2026 발표 예정
  TurboQuant가 그 결과를 달성하기 위해 사용하는 Quantized Johnson-Lindenstrauss(QJL) 및 PolarQuant(AISTATS 2026 발표 예정)도 공개

TurboQuant



자료: Google Research, 대신증권 Research Center

KV Cahce



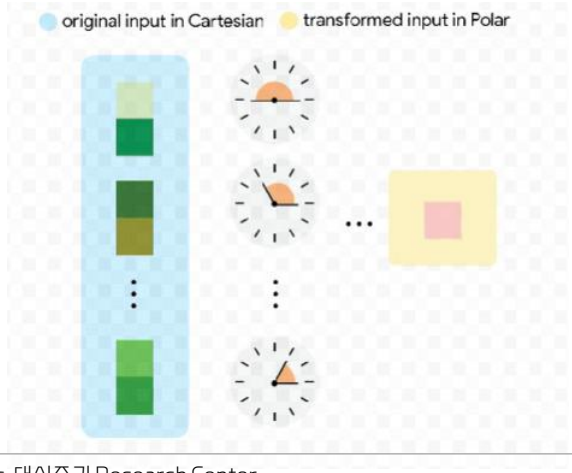
자료: NotebookLM, 대신증권 Research Center

1. AI 산업 트렌드

① 구글 리서치는 TurboQuant, PolarQuant, QJL 도입 → 정확도 손실 없이 KV 캐시 압축

- TurboQuant는 정확도 손실 없이 모델 크기를 대폭 줄여 KV 캐시 압축과 벡터 검색 모두를 지원하는 압축 방법. 두 가지 단계를 거침
- PolarQuant (고품질 압축): 표준 좌표계(X, Y, Z)를 사용하여 거리를 표시하는 대신, PolarQuant는 데카르트 좌표계를 사용하여 벡터를 극좌표로 변환
  이는 "동쪽으로 3블록, 북쪽으로 4블록 가세요"라는 명령을 "37도 각도로 총 5블록 가세요"로 바꾸는 것과 비슷
  결과적으로 반지름(핵심 데이터의 강도)과 각도(데이터의 방향 또는 의미)라는 두 가지 정보가 생성
  각도의 패턴은 알려져 있고 고도로 집중되어 있기 때문에, 모델은 경계가 끊임없이 변하는 "사각형" 대신 경계가 이미 알려진 고정된 예측 가능 "원형"에 매핑
  덕분에 PolarQuant는 전통적인 방식이 감수해야 했던 값비싼 데이터 정규화 단계와 메모리 오버헤드를 제거
- QJL 알고리즘 (숨겨진 오류 제거): TurboQuant는 아주 적은 잔여 압축력(단 1비트)을 사용하여 첫 번째 단계에서 남은 미세한 오류에 QJL 알고리즘을 적용
  QJL 단계는 편향을 제거하는 수학적 오류 체크 역할을 하여 더욱 정확한 어텐션 스코어를 도출
  QJL은 존슨-린덴슈트라우스 변환이라는 수학적 기법을 사용하여 데이터 포인트 간의 필수적인 거리와 관계를 보존하면서 복잡한 고차원 데이터를 축소
  각 결과 벡터 숫자를 단 하나의 부호 비트(+1 또는 -1)로 줄임. 이 알고리즘은 메모리 오버헤드가 전혀 없는 고속 속기(shorthand)를 만드는 것과 같음

PolarQuant



자료: Anthropic, 대신증권 Research Center

직교좌표계 → 극좌표계

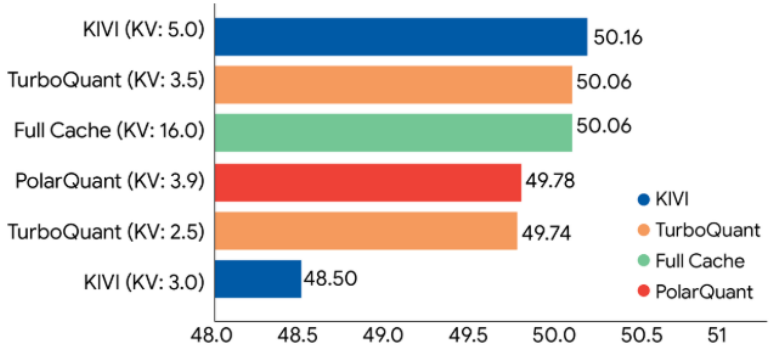
자료: NotebookLM, 대신증권 Research Center

1. AI 산업 트렌드

① 구글 리서치는 TurboQuant, PolarQuant, QJL 도입

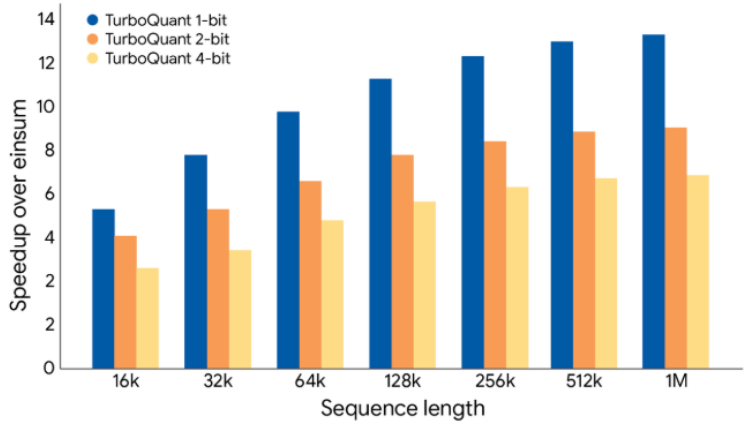
- 오픈 소스 LLM을 사용하여 LongBench, Needle In A Haystack, L-Eval 등을 포함한 표준 롱 컨텍스트 벤치마크에서 세 가지 알고리즘을 모두 평가
  실험 데이터에 따르면 TurboQuant는 내적 왜곡과 재현율 측면에서 최적의 스코어링 성능을 달성하는 동시에 KV 메모리 점유 공간을 최소화
- 아래 왼쪽 차트는 AI 모델 압축 기술의 성능을 비교한 것. 가로축은 모델의 정확도 점수, 오른쪽으로 갈수록 성능이 우수함을 의미
  세로축은 알고리즘 및 KV(비트 수)를 표현. 숫자가 작을수록 데이터를 더 많이 압축. 예) 16.0 = 원본 데이터, 2.5 = 원본 데이터를 6.4배 압축
  놀라운 점은 Full Cache(16.0)와 TruboQuant(3.5)의 점수가 50.06으로 동일하다는 점
  기존 기술 KIVI = 2024년 발표된 대표 저비용 양자화 알고리즘. KIVI(3.0)의 경우 점수가 48.5로 유의미하게 하락
- Long-context "needle-in-haystack" 작업(방대한 텍스트 속에 묻힌 특정 정보를 찾는 테스트) 결과
  TurboQuant는 모든 벤치마크에서 완벽한 다운스트림 결과를 달성하면서도 KV 메모리 크기를 최소 6배 줄임
- 아래 오른쪽 차트는 TurboQuant 기술을 적용했을 때 AI 모델의 연산 속도가 얼마나 빨라지는지를 보여주는 벤치마크 결과
  입력하는 텍스트의 양 = 시퀀스의 길이가 늘어날수록, 그리고 더 많이 압축할수록 속도 향상 폭이 커짐
  세로 축은 표준적인 연산 방식 대비 몇 배나 더 빨리 답을 도출하는지를 나타냄

압축 방식별 성능



자료: Google Research, 대신증권 Research Center

TurboQuant 양자화 수준에 따른 속도 향상 비교



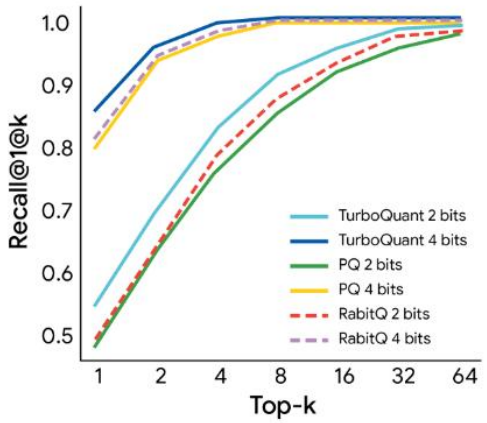
자료: Google Research, 대신증권 Research Center

1. AI 산업 트렌드

① 구글 리서치는 TurboQuant, PolarQuant, QJL 도입 → 메모리 효율적 사용 시도 지속되나 수요는 오히려 증가

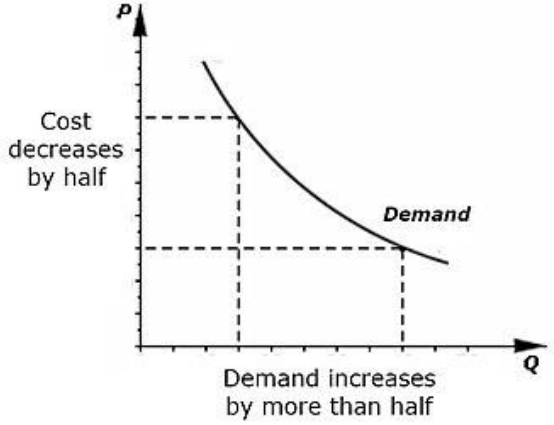
- 고차원 벡터 검색에서 TurboQuant는 최신 기술들 (PQ, RabitQ) 대비 더 정확하게 정보를 찾음 = 검색 재현율
- 왼쪽 아래 차트를 보면 모든 비트 설정(2비트, 4비트)에서 TurboQuant가 기존 방식인 PQ와 RabitQ보다 더 높은 재현율을 기록하며 가장 정확한 검색 결과 제공
  TurboQuant 4-bit(진한 파랑) 는 Top-4 지점에서 이미 재현율 1.0(100%)에 근접하며, 경쟁 모델들보다 훨씬 빠르게 정답을 찾아냄
  Top-k는 알고리즘이 추천한 상위 k개 결과 안에 실제 정답이 포함되어 있을 확률을 의미. 1.0에 가까울수록 정확
- 이는 대규모 데이터베이스에서 시맨틱 검색을 수행할 때, 서버 메모리는 1/8로 줄이면서도 검색 품질은 원본과 다름없는 수준으로 유지할 수 있다는 공학적 근거
- TurboQuant 발표 이후 메모리 기업들 주가 급락. 메모리를 효율적으로 사용할 경우 수요가 줄어들 수 있다는 우려 때문
  그러나 1) TurboQuant 구현 기술이 이미 1년 전에 발표됐다는 점, 2) 제번스의 역설로 효율적 사용이 실제 수요를 더 늘릴 것이라는 반박이 제기
- 명확한 것은 Agentic AI가 등장하면서 AI 연산 수요가 급증하고 있다는 점, 메모리 수요가 급증하자 이를 효율적으로 사용하려는 시도는 계속 이루어질 것이라는 점
- 메모리 수요는 지속 증가하겠으나 이를 효율적으로 사용하려는 수요는 지속적으로 나타날 것으로 예상

기술별 검색 재현율 비교



자료: Google Research, 대신증권 Research Center

제번스의 역설



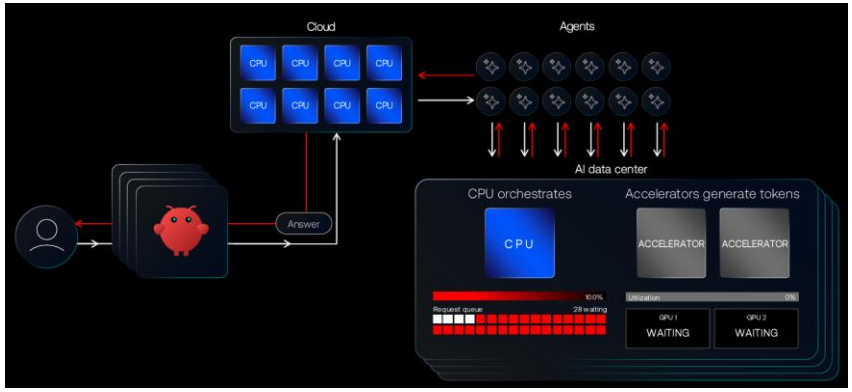
자료: Appropedia, 대신증권 Research Center

1. AI 산업 트렌드

② Arm Everywhere – Agentic AI 시대에서는 CPU가 새로운 병목이 될 것

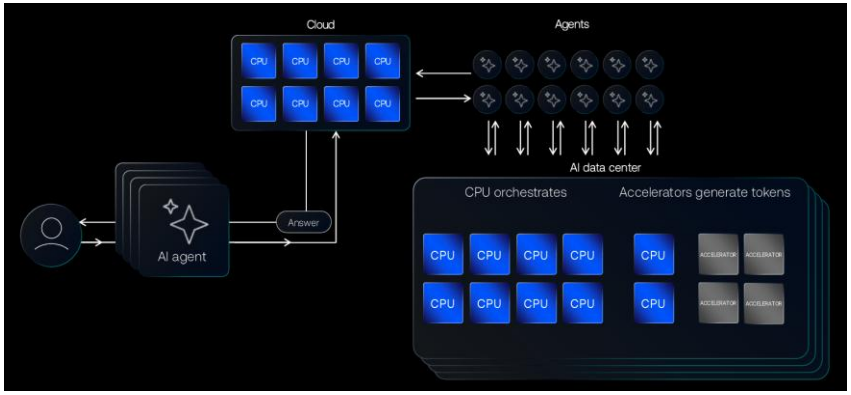
- Agentic AI 시대가 도래하면서 CPU 수요가 증가하기 시작
- Agentic AI 시대는 단순히 질문에 답을 하는 '채팅'의 단계를 넘어, AI가 스스로 목표를 세우고 도구를 사용하며 결과를 도출하는 '실행'의 단계
- 이 과정에서 전체 시스템을 조율하고 복잡한 명령을 내리는 CPU의 가치는 오히려 더 커지게 됨
- 첫째, 에이전트의 '사고 체계'는 순차적. GPU는 데이터를 병렬 계산하는데 강한 반면, CPU는 고성능 싱글 코어를 바탕으로 복잡한 조건문을 처리하는 데 최적화
  에이전트가 '생각의 고리(Chain of Thought)'를 이어가며 의사결정을 내리는 과정은 결국 CPU의 논리 연산 성능에 의존
- 둘째, 세상과의 연결 통로인 '도구 활용'은 CPU의 고유 영역.에이전트가 웹 서핑을 하고, 파이썬 코드를 실행하며, 이메일을 보내기 위해서는 운영체제를 써야 함
  파일 시스템에 접근하거나 네트워크 프로토콜을 처리하는 등의 시스템 호출(System Call)은 오직 CPU만이 관리할 수 있는 영역
  아무리 똑똑한 AI 모델(GPU)이라도, 실제 소프트웨어 환경을 조작해 일을 완수하려면 유능한 지휘관(CPU)의 손길이 필요

Agentic AI가 도입되면 CPU 병목이 생기는 이유



자료: Arm, 대신증권 Research Center

CPU 배치를 늘려 병목을 해소



자료: Arm, 대신증권 Research Center

1. AI 산업 트렌드

② Arm Everywhere – 4Q25 실적부터 CPU 수요 증가 확인

- CPU 수요가 증가하고 있다는 신호는 4Q25 실적 발표부터 나타나기 시작
- AMD – Agentic AI 및 신규 AI 워크로드가 헤드 노드 구동과 GPU와의 병렬 작업 처리를 위해 고성능 CPU를 필요로 함에 따라 EPYC CPU에 대한 수요가 급증
- 인텔 – 2025년 CPU TAM 전년대비 20% 성장. 2026년에는 CPU가 다시 쿨해지고 있음
- 인텔은 서버 CPU 수요에 생산 캐파를 더 할당하면서 데스크탑 CPU 가격을 인상하기 시작
  1분기에는 엔트리급 CPU 가격 15% 인상. 2~3분기에는 주력 및 중고가 CPU 가격도 인상할 예정
- ARM – FY 3Q26 데이터센터 매출은 전년대비 세 자릿수 성장세를 이어감
  향후 2년~3년 사이에 데이터센터 매출이 스마트폰 매출을 상회할 것으로 전망
  FY 3Q26 기준 데이터센터 매출은 20% 수준. 스마트폰 매출은 45% 수준으로 알려짐

인텔, AMD CPU 가격 15% 인상

Home > News > Components > Processors

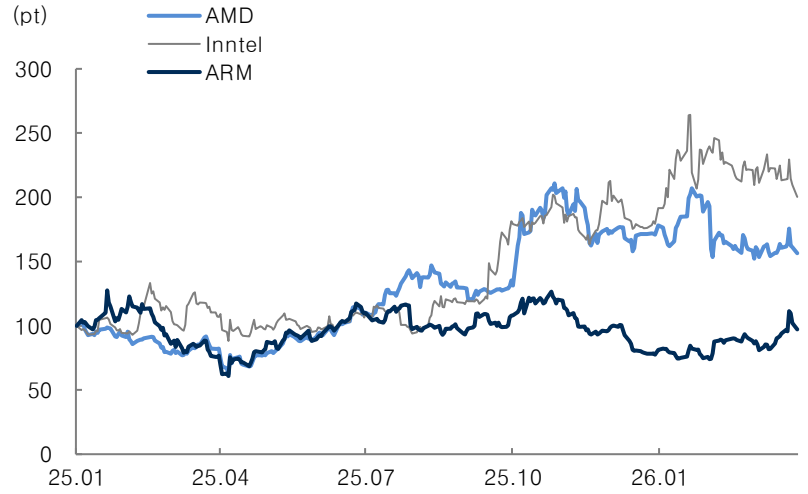
Intel, AMD Reportedly Set to Raise CPU Prices By Up to 15%

The increase is on tap even as some models have seen big price reductions recently.

By Jon Martindale March 25, 2026

자료: PCMag, 대신증권 Research Center

CPU 관련 기업들 주가



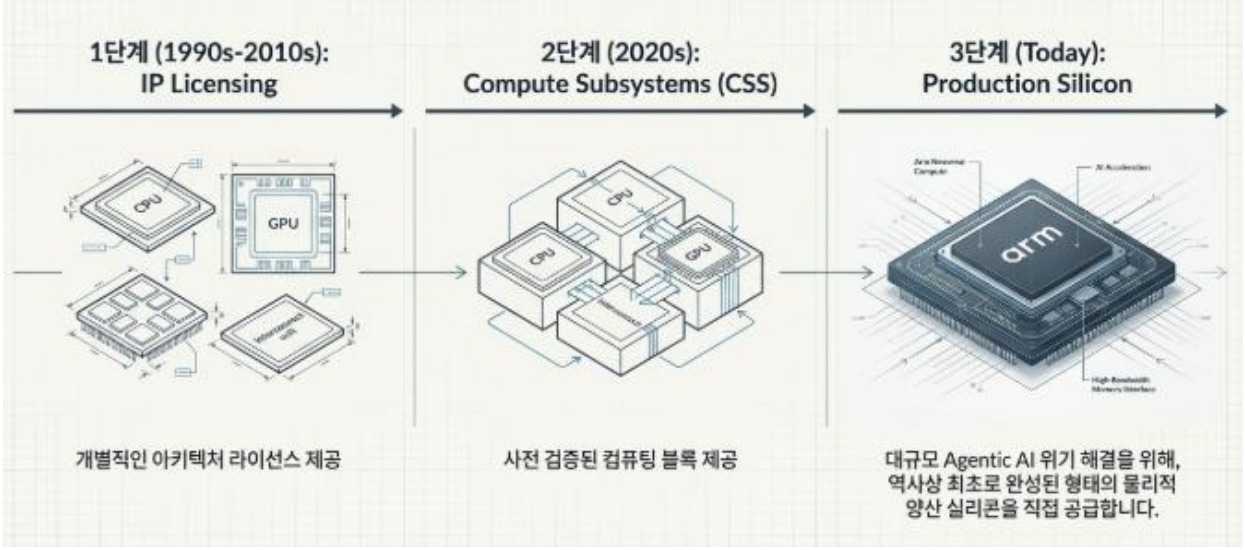
자료: Factset, 대신증권 Research Center

1. AI 산업 트렌드

ⓔ Arm Everywhere – Arm Everywhere 행사를 통해 제시된 장기 전략

- 3월 25일 ARM 경영진은 Arm Everywhere 행사를 통해 장기 전략을 제시
- Arm은 1990년대~2010년대 반도체 설계 IP를 라이선스하는 사업으로 시작
  이후 2020년대는 Compute Subsystems (CSS)로 불리는 사전 통합 및 검증된 컴퓨팅 블록을 제공 = 초미세 공정에서 설계 비용과 시간을 크게 절감
  Arm Everywhere 행사에서는 IP 라이선스 뿐만 아니라 자체 설계한 서버 CPU를 판매하는 방향으로 전략을 변경
- Arm AGI CPU는 2026년 하반기부터 출하가 시작될 전망
- 메타가 리드 파트너로 참여하며 처음으로 대규모 도입할 예정. OpenAI, Cloudflare, Cerebras, SAP 등도 주요 파트너사로 이름을 올림
- 메타는 Arm 기반의 독립형 엔비디아 Grace CPU 서버도 도입하는 계약을 체결한 바 있음

ARM 방향 변화



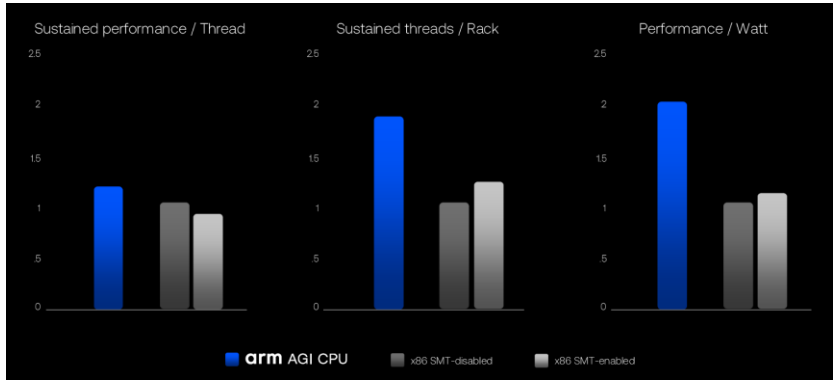
자료: Arm, 대신증권 Research Center

1. AI 산업 트렌드

② Arm Everywhere – Arm 자체 설계 칩인 AGI CPU 공개

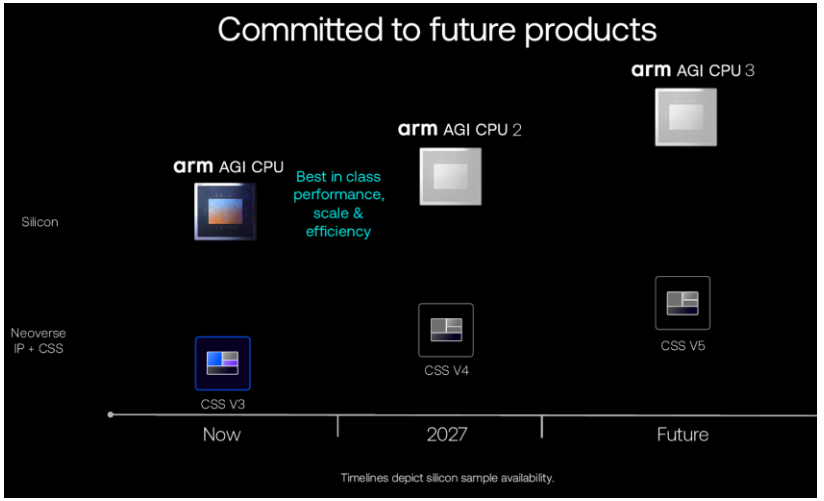
- Arm 아키텍처 기반 CPU는 AMD, 인텔의 x86 아키텍처 기반 CPU 대비 저전력이라는 특성을 바탕으로 모바일폰 시장을 장악
- 이러한 특징을 기반으로 Arm은 서버 CPU 부문으로도 확장. Arm은 하이퍼스케일러 기준으로 점유율 50%를 상회하고 있다고 밝힘
- 1GW AI 데이터센터 건설을 기준으로 Arm CPU를 사용할 경우 x86 CPU 대비 100억 달러를 절감할 수 있다고 주장
- 이는 Arm CPU가 x86 CPU 대비 와트당 성능이 2배 수준을 기록하는 점에 기인
- x86 CPU가 왼쪽 아래 차트처럼 SMT(Simultaneous Multi-Threading) 방식을 사용하더라도 이와 같은 성능 차이가 유지됨
- Arm은 이를 바탕으로 2031년 칩 판매 매출이 150억 달러를 상회할 것으로 전망
  아울러, 2030년 기준 Arm의 TAM은 1조 달러에 달할 것으로 예상

Arm AGI CPU vs x86 CPU 성능 비교



자료: Arm, 대신증권 Research Center

Arm AGI CPU 로드맵



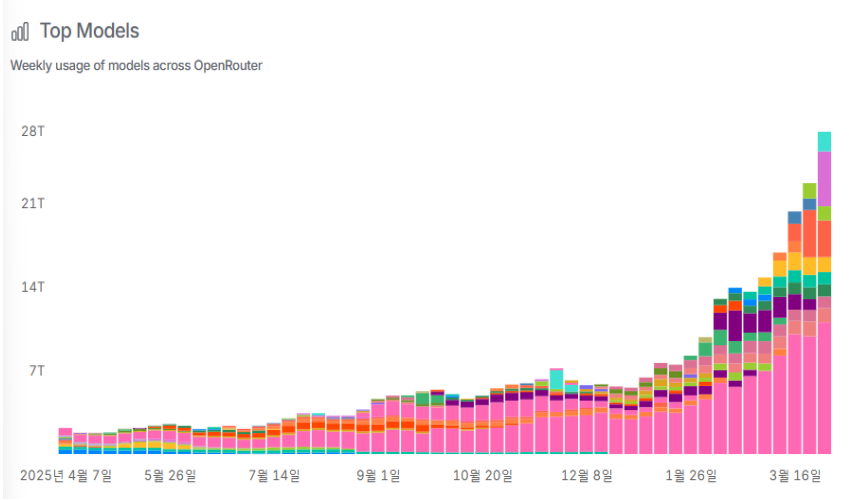
자료: Arm, 대신증권 Research Center

1. AI 산업 트렌드

③ Agentic AI 수요 증가 → 주간 토큰 생성량의 폭발적인 증가

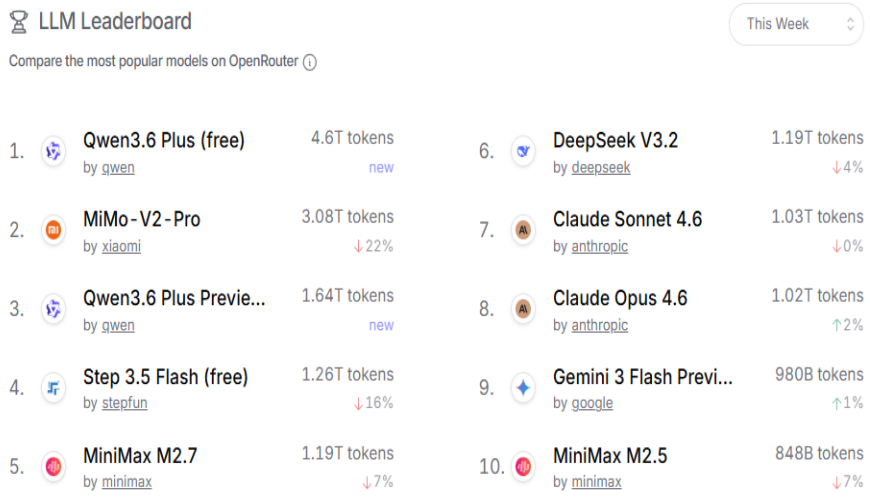
- 지난해 말 발표된 Claude Opus 4.5를 기점으로 Agentic AI 시대 도래
- OpenRouter의 주간 토큰 생성량을 보면 지난해 말부터 급격하게 증가. 토큰 생성량은 지난해 말 대비 4배 가까이 증가한 것으로 추정
- 토큰 생성량 증가를 보면 클라우드 기업들의 CapEx는 더욱 증가할 가능성이 높다고 판단
- 주요 클라우드 기업들의 실적 발표가 2월 초에 이루어진 점을 고려하면 지금의 토큰 생성량이 반영되지 않았을 가능성 존재
- LLM Leaderboard를 보면 중국의 LLM들이 상위권을 차지
- 이는 쉽고, 간단한 태스크는 API 가격이 낮은 중국 모델을 쓰는 결과
- 어렵고 복잡한 태스크의 경우에는 클로드, 제미나이, 그록을 쓰는 모습을 보임

OpenRouter 주간 토큰 생성량



자료: OpenRouter, 대신증권 Research Center

OpenRouter LLM 리더보드



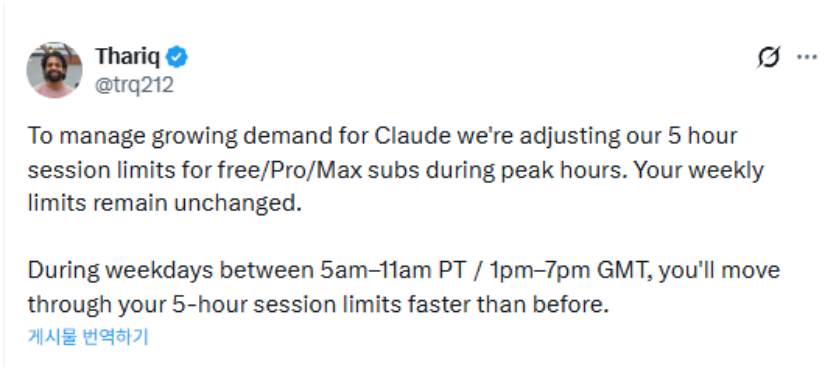
자료: OpenRouter, 대신증권 Research Center

1. AI 산업 트렌드

③ Agentic AI 수요 증가 → 컴퓨팅 캐파 제약으로 사용량을 제한하는 AI 기업들

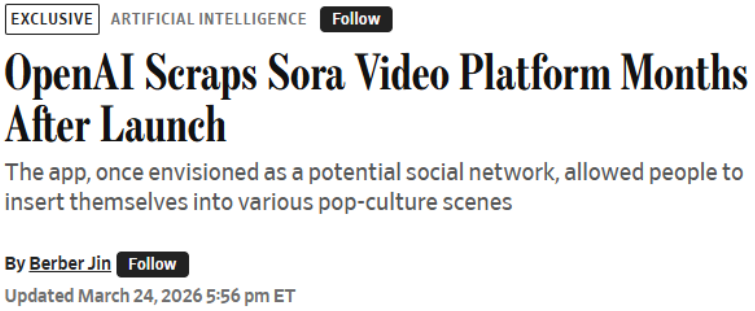
- Agentic AI 수요 증가에 따른 컴퓨팅 캐파의 제약을 반영하듯 주요 AI 기업들은 사용량을 제한하거나, 자원을 집중하는 경향을 보이기 시작
- 앤트로픽은 컴퓨팅 수요 급증을 감당하기 위해 피크 시간대 클로드 구독 요금제 사용자의 세션 제한을 강화
- 평일 오전 5시~오전 11시(PT 기준) / 오후 1시~오후 7시(GMT 기준) 사이에는 이전보다 5시간 세션 제한에 더 빨리 도달
- OpenAI는 앤트로픽과 알파벳과의 경쟁이 치열해짐에 따라 컴퓨팅 자원을 기업용 제품 개발에 더 투입하기로 결정
- 이에, 컴퓨팅 자원을 막대하게 소모하면서도 수익 기여는 상대적으로 적었던 동영상 생성 AI 플랫폼 Sora를 폐쇄하기로 결정
- Sora는 하루에 백만 달러의 손실을 낸 것으로 알려짐. 콘텐츠 관리 문제와 저작권 노출에 대한 우려가 커진 점도 원인으로 작용
- 이를 반영하듯 A100, H100, B200 시간당 렌탈 가격은 3월 지속적인 상승세를 보임

앤트로픽은 피크 타임 세션 제한 강화



자료: X, 대신증권 Research Center

OpenAI – Sora 플랫폼을 폐쇄하기로 결정



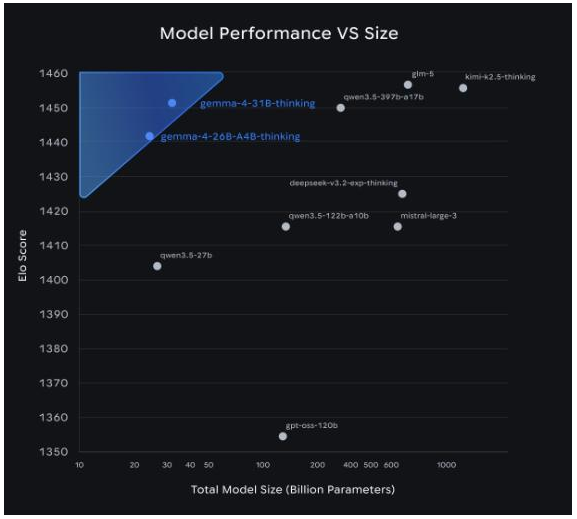
자료: WSJ, 대신증권 Research Center

1. AI 산업 트렌드

④ 알파벳의 Gemma 4 → 온디바이스 AI 신호탄

- 구글 딥마인드가 고급 추론과 에이전트 워크플로우에 최적화된 차세대 오픈 모델 Gemma 4를 공개
- Gemini 3와 동일한 연구 및 기술을 기반으로 구축된 Gemma 4는 클라우드가 아닌 사용자가 보유한 하드웨어에서 실행할 수 있는 가장 뛰어난 모델
- Gemma 4는 네 가지 크기로 출시: Effective 2B (E2B), Effective 4B (E4B), 26B Mixture of Experts (MoE), 그리고 31B Dense
- Gemma 4는 자신보다 20배 큰 모델들과 대등하게 경쟁. 31B 모델은 현재 업계 표준인 Arena AI 텍스트 리더보드에서 세계 3위 오픈 모델로 랭크
- 수십억 대의 안드로이드 기기부터 노트북 GPU, 개발자 워크스테이션 및 가속기에 이르기까지 하드웨어에서 효율적으로 실행하고, 조종하도록 설계
- 아래 왼쪽 차트를 보면 Open Weight 모델 경쟁자인 qwen, deepseek, llama 대비 파라미터는 가장 적은 반면, 성능은 가장 높음

모델 사이즈 대비 성능 측면에서 압도적인 우위



자료: Google DeepMind, 대신증권 Research Center

Gemma 4 성능

Benchmark		Gemma 4 31B IT Thinking	Gemma 4 26B A4B IT Thinking	Gemma 4 E4B IT Thinking	Gemma 4 E2B IT Thinking
Arena AI (text) As of 4/2/26		1452	1441	-	-
MMMLU Multilingual Q&A	No tools	85.2%	82.6%	69.4%	60.0%
MMMU Pro Multimodal reasoning		76.9%	73.8%	52.6%	44.2%
AIME 2026 Mathematics	No tools	89.2%	88.3%	42.5%	37.5%
LiveCodeBench v6 Competitive coding problems		80.0%	77.1%	52.0%	44.0%
GPQA Diamond Scientific knowledge	No tools	84.3%	82.3%	58.6%	43.4%
τ2-bench Agentic tool use	Retail	86.4%	85.5%	57.5%	29.4%

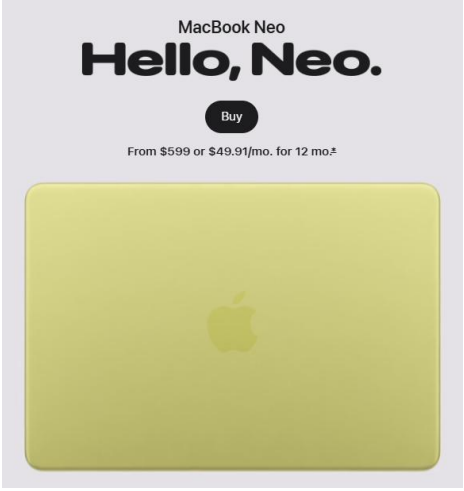
자료: Google DeepMind, 대신증권 Research Center

1. AI 산업 트렌드

④ 알파벳의 Gemma 4 → 점유율 확대를 노리는 애플

- 애플은 메모리 및 저장장치 가격의 급등에도 불구하고 신형 보급형 모델인 iPhone 17e와 MacBook Neo의 가격을 599달러로 책정하며 시장 점유율 확대 추진
- 원가 상승 압박으로 인해 가격을 올리거나 수익성을 포기해야 하는 중저가 안드로이드 및 PC 제조사들의 약점을 공략해 사용자들을 유인하려는 전략
- 보급형 제품에서 발생하는 마진 손실은 MacBook Pro와 같은 프리미엄 제품군의 가격 인상을 통해 상쇄
- 보급형 모델인 iPhone 17e의 시작가는 작년 모델과 동일한 599달러로 정해짐. 보급형 노트북인 MacBook Neo를 예상보다 낮은 599달러에 발표
- 애플이 막대한 자금력을 바탕으로 시장의 가용 모바일 DRAM을 고가에 싹쓸이하며, 경쟁사들이 메모리 칩을 확보하지 못하게 만드는 공격적인 '공급망 차단' 전략
- 중저가 스마트폰에 주로 사용되는 4nm 칩의 생산량을 미디어텍(MediaTek)과 퀄컴(Qualcomm) 모두 감축
- 이러한 생산 축소 규모는 약 20,000~30,000개의 웨이퍼에 달하며, 이는 약 1,500만~2,000만 개의 모바일 칩 분량

MacBook Neo



자료: Apple, 대신증권 Research Center

모바일폰 메모리를 사들이는 애플

자료: wccftech, 대신증권 Research Center

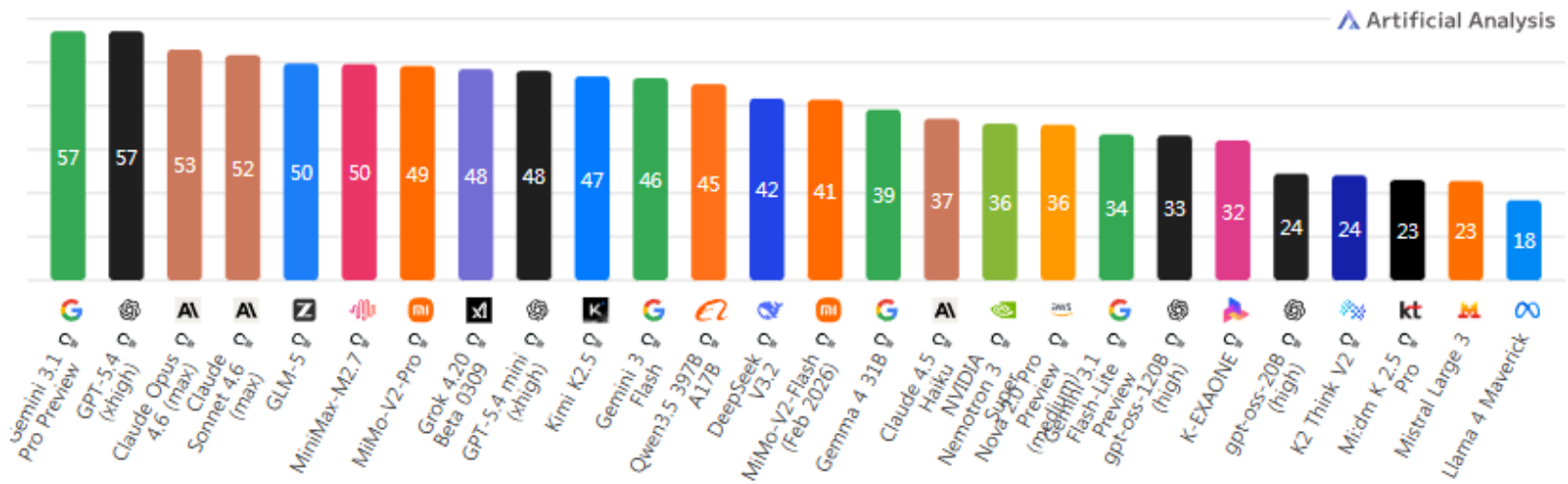
II. AI 뉴스 & 데이터 대시보드

3. AI 데이터 대시보드

대형언어모델 성능 리더보드

- Artificial Analysis의 최신 LLM 리더보드에 따르면,
  1위: Gemini 3.1 Pro Preview (57점), 2위: GPT-5.4 (57점), 3위: Claude Opus 4.6 (53점)
- 앤트로픽이 Claude Opus 4.6 (Max)를 통해 GPT-5.3 Codex를 밀어내고 3위로 한 단계 상승
- 주목할 점은 중국 기업들의 순위 상승.
- AI 기업 Z.ai가 출시한 GLM-5의 경우 클로드의 뒤를 이어 5위를 차지
- 미니맥스가 만든 M2.7, 샤오미가 만든 MiMo-V2-Pro의 경우도 각각 6위, 7위를 차지

Artificial Analysis 대형언어모델 리더보드



주: 인덱스 점수는 MMLU-Pro, GPQA Diamond, Humanity's Last Exam, LiveCodeBench, SciCode, AIME, MATH-500 평가를 기반으로 함
자료: Artificial Analysis, 대신증권 Research Center

3. AI 데이터 대시보드

국내 상장 AI 소프트웨어 ETF 리스트

티커	ETF명	기초지수명	AUM (억원)	총보수 (%)	상장일	수익률(%)					
						1주	1개월	3개월	6개월	12개월	연초 대비
A456600	TIMEFOLIO 글로벌AI 인공지능액티브	Solactive Global Artificial Intelligence 지수	11,909	0.80	2023.05.16	2.0	3.0	21.8	17.2	95.5	22.6
A466950	TIGER 글로벌AI액티브	Indxx Artificial Intelligence and Big Data 지수(PR)	3,467	0.79	2023.10.11	3.4	0.8	11.6	9.5	72.4	11.6
A481180	SOL 미국AI소프트웨어	KEDI 미국 AI 소프트웨어지수(PR)	1,425	0.45	2024.05.14	-0.3	1.1	-19.6	-23.7	1.0	-19.9
A0040Y0	SOL 팔란티어커버드콜OTM 채권혼합	KEDI 팔란티어커버드콜OTM 채권혼합지수(NTR)	3,520	0.35	2025.04.22	-0.8	2.0	-3.6	-0.1	-	-4.0
A0102A0	TIGER 미국AI소프트웨어TOP4Plus	Akros U.S. AI Software TOP4 Plus 지수 (Price Return)	822	0.49	2025.09.09	-2.2	-4.1	-27.7	-36.4	-	-28.3
A412770	TIGER 글로벌AI플랫폼액티브	Indxx Global Metaverse Index (Price Return)	860	0.79	2021.12.22	-0.7	4.0	-12.0	-19.9	10.6	-11.1
A0040X0	SOL 팔란티어미국채커버드콜혼합	KEDI 팔란티어미국채커버드콜혼합지수(NTR)	2,195	0.35	2025.04.22	0.1	2.8	-1.1	1.0	-	-1.7
A0041D0	KODEX 미국AI소프트웨어TOP10	iSelect 미국 AI 소프트웨어 TOP10 지수(PR)	613	0.45	2025.04.22	0.3	1.6	-22.1	-26.0	-	-22.4
A0093D0	KoAct 팔란티어밸류체인 액티브	KEDI 팔란티어밸류체인지수(PR)	306	0.50	2025.08.26	2.8	1.5	15.6	23.0	-	15.6
A0047R0	RISE 팔란티어고정테크100	KEDI 팔란티어고정테크100지수(PR)	345	0.20	2025.05.13	1.2	1.4	-6.2	-1.4	-	-6.2
A0067Y0	TIGER 차이나AI소프트웨어	Mirae Asset China AI Software 지수 (Price Return, HKD)	172	0.49	2025.06.17	0.0	-0.9	-4.5	-12.1	-	-0.6
A0015F0	KIWOOM 팔란티어미국30년국채혼합액티브(H)	Bloomberg Blended PALANTIR and US Long Treasury Bond Index	177	0.19	2025.02.11	0.8	-1.1	-4.3	-6.3	11.1	-4.9
A0105D0	SOL 한국AI소프트웨어	KEDI 한국AI소프트웨어 지수(PR)	134	0.45	2025.09.23	-5.6	-16.7	-14.0	-7.9	-	-12.1
A0050E0	PLUS 미국AI에이전트	Solactive US AI Agents Index(PR)	38	0.45	2025.05.20	0.0	-0.2	-22.2	-20.8	-	-22.4

자료: Quantiwise, 대신증권 Research Center

3. AI 데이터 대시보드

국내 상장 AI 인프라 ETF 리스트

티커	ETF명	기초지수명	AUM (억원)	총보수 (%)	상장일	수익률(%)					
						1주	1개월	3개월	6개월	12개월	연초 대비
A487230	KODEX 미국AI전력핵심인프라	iSelect 미국AI전력핵심인프라 지수(Price Return)	14,069	0.45	2024.07.09	0.8	0.2	21.9	11.8	92.5	23.6
A487240	KODEX AI전력핵심설비	iSelect AI 전력핵심설비 지수(Price Return)	18,266	0.39	2024.07.09	-3.0	-3.0	29.0	78.2	257.7	35.4
A486450	SOL 미국AI전력인프라	KEDI 미국AI전력인프라 지수(PR)	2,495	0.45	2024.07.16	-0.3	-1.1	9.9	7.1	80.3	10.6
A491010	TIGER 글로벌AI전력인프라액티브	Mirae Asset AI Infrastructure 지수(PR)	5,130	0.49	2024.09.10	1.8	-0.1	29.0	25.2	120.7	29.4
A486240	DAISHIN343 AI반도체&인프라액티브	FnGuide AI반도체&인프라 지수	644	0.36	2024.06.18	-2.4	-2.7	34.3	71.1	172.7	41.8
A487130	KoAct AI인프라액티브	Solactive AI인프라 지수	313	0.50	2024.07.09	-3.5	-1.6	30.8	57.2	177.3	37.3
A489860	KIWOOM 글로벌전력GRID인프라	NASDAQ OMX Clean Edge Smart Grid Infrastructure Index	98	0.49	2024.08.27	0.8	-3.4	11.5	15.6	52.5	11.6
A489010	PLUS 글로벌AI인프라	Solactive Global AI Infrastructure Index (PR)	27	0.45	2024.08.13	0.5	0.2	13.3	12.0	67.8	13.7

자료: Quantiwise, 대신증권 Research Center

Compliance Notice

- 금융투자업규정 4-20조 1항5호사목에 따라 작성일 현재 사전고지와 관련한 사항이 없습니다.
- 당사의 금융투자분석사는 자료 작성일 현재 본 자료에 관련하여 재산적 이해관계가 없습니다. 당사는 위 언급된 종목을 제외한 동자료에 언급된 종목과 계열회사의 관계가 없으며 당사의 금융투자분석사는 본 자료의 작성과 관련하여 외부의 부당한 압력이나 간섭을 받지 않고 본인의 의견을 정확하게 반영하였습니다. (작성자: 서영재)
- 본 자료는 투자자들의 투자판단에 참고가 되는 정보제공을 목적으로 배포되는 자료입니다. 본 자료에 수록된 내용은 당사 Research Center의 추정치로서 오차가 발생할 수 있으며 정확성이나 완벽성은 보장하지 않습니다. 본 자료를 이용하시는 분은 동 자료와 관련한 투자의 최종 결정은 자신의 판단으로 하시기 바랍니다.