Research Center | 2025.12.09

Daishin 증권

AI Monthly

구글의 반격, Gemini 3.0

▪ 박세라 sera.park@daishin.com

Contents

I. AI 산업 트렌드 05

II. AI 기술 동향 11

III. AI 뉴스 & 데이터 대시보드 18

Summary

구글의 반격, Gemini 3.0

- 구글은 Gemini 3.0을 통해 "Thinking Model + Agentic AI + Native Multimodality"를 결합한 고도화된 멀티모달 AI를 공개했으며, **칩(TPU)-클라우드 인프라-운영 플랫폼-자체 프론티어 모델까지 수직 계열화 전략과 동시에 엔드 투 엔드 AI 생태계를 보유하게 되었다.**
- Gemini 3.0은 컨텍스트 윈도우를 대폭 확장해 수백 권 분량 문헌과 장시간 영상을 한 번에 처리할 수 있다. Gemini 3.0의 Deep Think 모드는 Humanity's Last Exam, GPQA Diamond, ARC-AGI-2 등 주요 벤치마크에서 높은 점수를 기록하면서, 복잡한 추론·검증·의사결정 지원 능력을 입증했다. 과거 전문가 고유 영역이었던 법률 검토, 금융 리스크 분석, 대규모 리서치 등 지식 노동의 자동화 범위가 빠르게 넓어지고 있으며, 장기 기억 기능을 바탕으로 사용자별 맥락과 선호를 축적하는 맞춤형 비즈니스 솔루션 구현도 가능해지고 있다.
- 구글은 자체 칩인 TPU 6세대 Trillium을 기반으로 전세대 대비 컴퓨팅 성능 4.7배, 에너지 효율 67% 개선을 달성하고, AI Hypercomputer 아키텍처로 수만 개 TPU를 하나의 초거대 슈퍼컴퓨터처럼 운용함으로써 학습·추론 비용 절감과 지연 시간 최소화를 동시에 추구한다. 이는 엔비디아 GPU와 외부 클라우드에 의존하는 오픈AI, 멀티클라우드 전략을 택한 앤트로픽과 대비되는 구글만의 비용우위·규모의 경제 전략으로 평가된다.
- 동시에 구글은 AI Overviews, AI 모드, Gemini 기반 검색에 쿼리 팬 아웃(Query Fan-out)과 MuVERA를 도입하며 검색 패러다임 전환을 주도하고 있다. 한 개의 질문을 5~10개 하위 쿼리로 확장해 주제별로 심층 리서치를 수행하는 쿼리 팬 아웃은 사용 경험과 만족도를 크게 높이지만, 일반 검색 대비 10~30배 수준의 비용 상승을 초래한다. 이를 상쇄하기 위해 구글은 대략적 필터링–핵심 정보 선별–지연 상호작용 3단계 구조의 MuVERA를 통해 불필요 연산의 90%를 제거하면서도 검색 정확도를 유지하는 경량화 전략을 병행하고 있다.
- **Gemini 3.0의 출시는 단순히 새로운 모델의 추가가 아니라, 모델–인프라–검색–플랫폼–업무툴이 연결된 '구글식 통합 AI 스택'의 본격 가동을 의미한다**. 구글의 반격은 AI 검색, 클라우드, 모바일·유튜브 생태계 전반에서 멀티모달·에이전트형 AI를 전면 배치하는 전략으로 나타나고 있으며, 이는 향후 AI 산업의 경쟁 기준을 "모델 성능"에서 엔드투엔드 운영체제로서의 완결성과 경제성 전환시키는 촉매로 작용할 것으로 예상한다.

I. 이달의 AI 산업 트렌드: 구글의 반격, Gemini 3.0

① 구글, Gemini 3.0 출시

② 멀티모달을 넘어선 '깊은 추론(Deep Reasoning)' 능력

③ Big 3 구글, 오픈 AI, 앤트로픽의 경쟁

④ Gemini 3.0을 지탱하는 강력한 칩 TPU

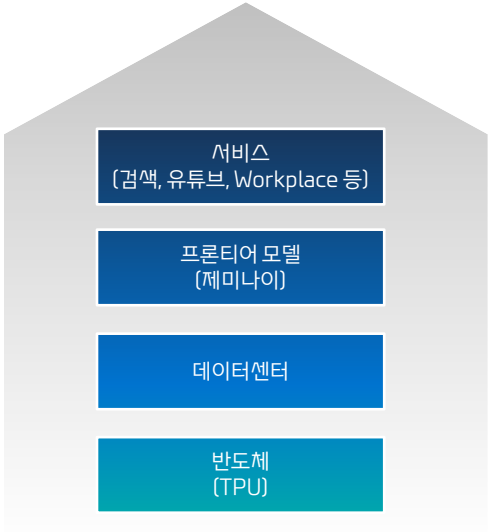
⑤ 구글 생태계와 AI의 시너지 효과

1. AI 산업 트렌드

①구글, Gemini 3.0 출시

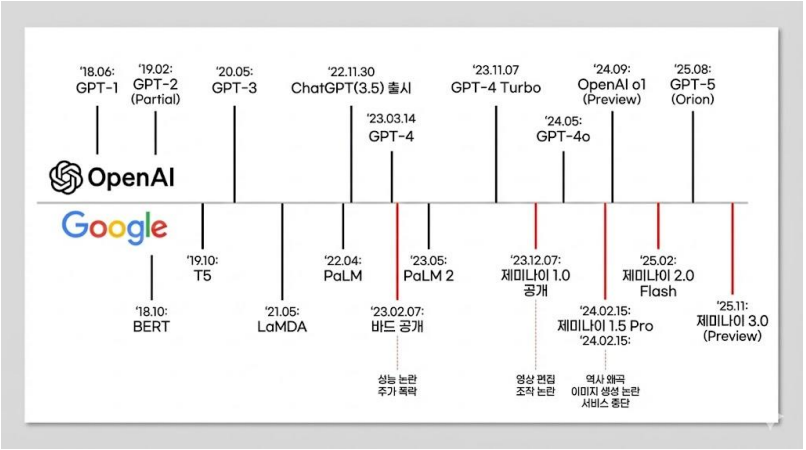
- 지난달 구글은 "Thnking Model + Agentic AI + Native Multimodality"라는 세가지 축을 통합한 고도화된 멀티모달 AI인 Gemini 3.0을 공개함
- Gemini 3.0은 자사 맞춤 하드웨어인 TPU(Tensor Processing Unit)를 사용하여 개발되었으며 이는 수직 계열화 전략으로, 엔비디아 GPU 의존을 벗어나 구글 생태계 전체를 고성능 AI 플랫폼으로 통합하려는 전략을 보여줌
- Gemini 3.0은 단순한 언어모델을 넘어서, 복잡한 추론과 도구호출 및 에이전트형 워크플로우(문서분석, 자동 코딩 등)를 가능하게 하며, 엔터프라이즈용 비즈니스 운영체제(OS)로서의 변화를 예고함
- <u>구글은 칩(TPU)-클라우드 인프라-운영 플랫폼-자체 프론티어 모델까지 수직 계열화 전략과 동시에 엔드 투 엔드 AI 생태계를 보유하게됨</u>

구글의 AI 기술 수직 계열화



자료: 대신증권 Research Center

Gemini 1 → 1.5 → 2.0 → 2.5 → 3.0 출시 타임라인



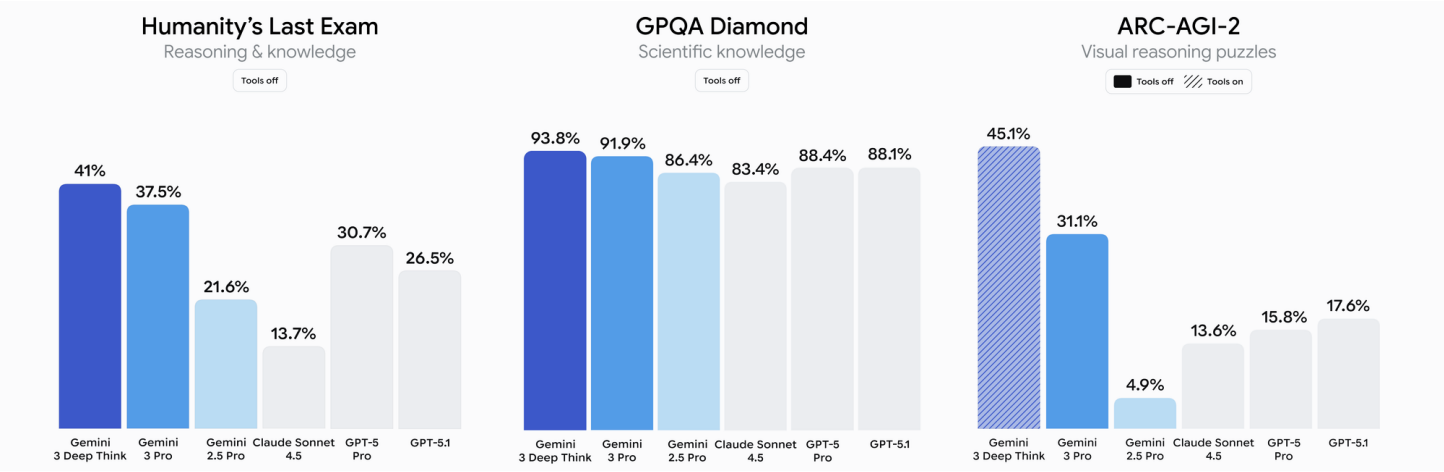
자료: 뉴스 자료 취합, 대신증권 Research Center

1. AI 산업 트렌드

②멀티모달을 넘어선 '깊은 추론(Deep Reasoning)' 능력

- Gemini 3.0은 이전 모델 대비 2배 늘어난 컨텍스트 윈도우로 수백 권 분량의 책 데이터와 긴 영상을 한 번에 처리함
- 특히, Deep Think 모드를 통해 복잡한 추론과 단계적 사고가 가능해지면서, AI가 단순 응답을 넘어 의사결정·분석·검증을 수행하는 패러다임으로 진입
- Deep Think 모드는 Humanity's Last Exam(41.0%), GPQA Diamond(93.8%), ARC-AGI-2(45.1%) 등 높은 벤츠마크 결과를 기록하면서 추론 성능을 입증함
- 과거에는 대규모 문헌 분석, 법률 검토, 금융 리스크 산출 등 전문 업무는 숙련된 전문가만이 가능한 영역이었지만, Gemini 3.0은 수백 권 분량의 데이터와 긴 영상을 한 번에 처리 가능하기 때문에 지식 노동의 자동화 범위를 확장 시키고 있음
- 또한, 장기 기억(Long-term Memory)가 가능하기에 사용자의 이전 상호작용과 선호도를 장기적으로 기억해 맞춤형 비즈니스 솔루션 제공도 가능

Gemini 3.0 Deep Think Mode는 가장 높은 벤치마크 결과 기록



자료: 구글, 대신증권 Research Center

1. AI 산업 트렌드

③Big 3 구글, 오픈 AI, 앤트로픽의 경쟁

- 현재 미국 내 AI는 구글, 오픈 AI, 앤트로픽의 AI 빅 3 구도로 재편되고 있으며, 각자의 생태계 및 기술 우위를 기반으로 AI 플랫폼 경쟁을 본격화 하고 있음
- 구글은 Gemini 3 + TPU + Google Cloud + Workspace 및 플랫폼(유튜브 등)를 하나의 통합 스택으로 제공해 AI 프론티어 모델– 인프라 – 업무툴 간 연결을 극대화 하고 있음
- 오픈 AI는 ChatGPT를 중심으로 GPT-5.1+ GPU +Azure+Sora 2 등으로 대규모 사용자 풀을 확장하고 있음. 앤트로픽의 경우 Claude 4.5와 AWS 및 구글 클라우드 파트너십을 통한 유연한 멀티 클라우드 대응을 펼치고 있으며 철저하게 B2B 시장을 공략하고 있음
- 경제성 측면에서는 구글이 칩(TPU)부터 서비스까지 수직 계열화 했기 때문에 훨씬 경제적임. 오픈 AI의 경우 마이크로소프트의 클라우드를 임대, 엔비디아의 칩을 구매해서 쓰는 마진 스태킹 구조이지만, 앤트로픽은 아마존으로부터 칩(Trainium)과 클라우드를 제공 받기 때문에 훨씬 경제적임

구글, 오픈 AI, 앤트로픽 최신 AI 모델 및 기술적 역량 비교

항목	구글 Gemini 3.0 (Pro/Ultra)	오픈 AI GPT-5 (Instant/Pro)	앤트로픽 Claude 4.5 (Opus)
핵심강점	**멀티모달 통합** 비디오/오디오 네이티브 멀티모달	**대화 및 창의성** 자연스러운 대화 및 멀티모달 기능	**신뢰성 및 코딩** 긴 문맥 처리 및 코딩 우수
추론 능력	**Deep Think** 추상적 수학/과학 문제 해결 최상위	**Thinking Mode** 복합적인 언어적 추론과 뉘앙스 파악 우수	**엄격한 논리** 할루시네이션이 적고 지시 이행력 우수
에이전트 방식	**생성형 (Generative UI)** 새로운 도구를 만들어 사용자에게 제공	**오케스트레이터 (Orchestrator)** 플러그인을 통한 외부 도구 호출	**Computer Use** 인간처럼 기존 화면을 직접 조작
인프라	**TPU v6** 자체 칩으로 비용 통제 및 데이터 센터 최적화	**Azure + NVIDIA:** 고성능이지만 높은 비용 구조 및 인프라 구축 지연	**AWS Trainium 2** 대규모 클러스터 및 비용 효율 확보

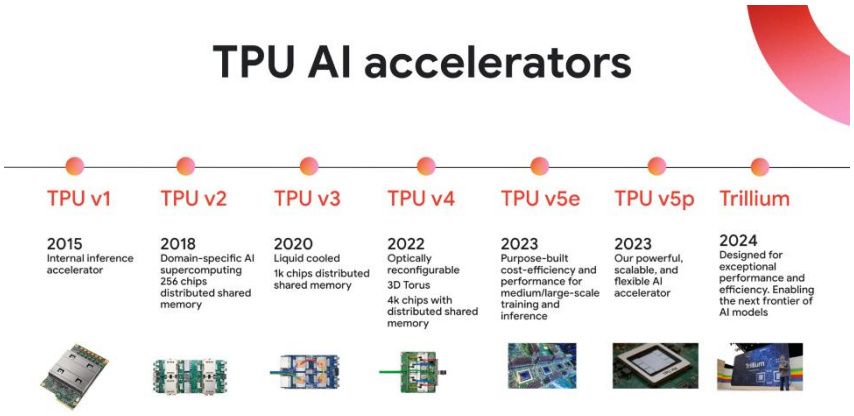
자료: 각 사, 대신증권 Research Center

1. AI 산업 트렌드

④Gemini 3.0을 지탱하는 강력한 칩 TPU

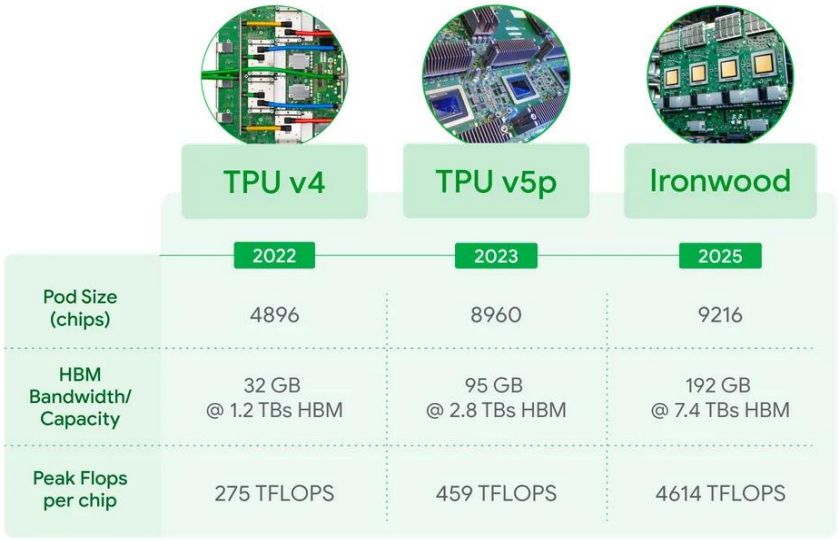
- 구글은 TPU 6세대 트릴리움(Trillium) 도입을 통해 전세대 대비 컴퓨팅 성능 4.7배 향상과 에너지 효율을 67% 개선하면서 AI 인프라 경쟁에서도 기술적 우위를 확보. 최신 칩인 Ironwood(TPU 7세대)는 Gemini 3.0의 추론 및 서빙 인프라로 투입될 예정
- 구글은 AI Hypercomputer 아키텍처를 통해 수만 개의 TPU를 초고속 인터커넥트로 연결해 하나의 초거대 슈퍼컴퓨터처럼 작동 시키는데 성공함
- 구글은 수직 계열화를 통해(칩 설계 → 데이터센터 → 클라우드 → AI 모델 최적화) 지연 시간 최소화, 추론 비용 절감, 초거대 모델 학습 시간 및 비용 단축을 동시에 달성함
- 특히 오픈 AI나 앤트로픽 등 엔비디아의 GPU 의존도가 높은 반면, 구글은 자체 칩 기반 운영으로 API 단가 경쟁력을 확보했기 때문에 시장에서 비용 우위 전략을 전개할 수 있음

TPU 시리즈 개발 타임라인



자료: 구글 클라우드, 대신증권 Research Center

최신 7세대 TPU인 Ironwood는 v5p 대비 약 10배 높은 성능



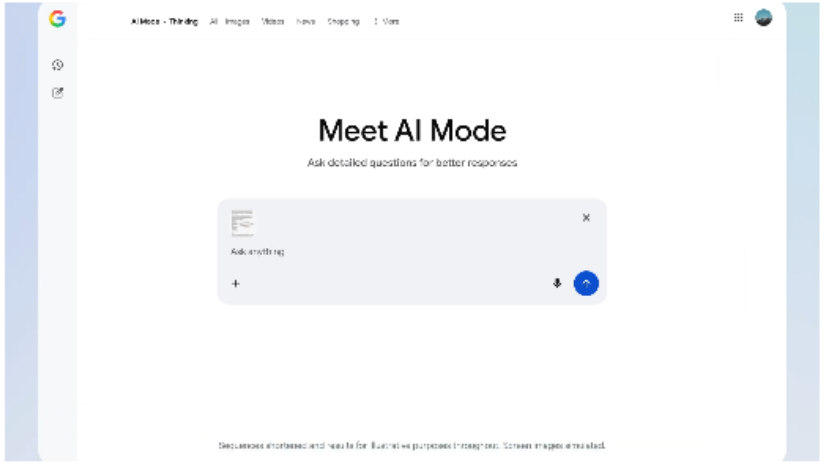
자료: 구글 클라우드, 대신증권 Research Center

1. AI 산업 트렌드

⑤구글 생태계와 AI의 시너지 효과

- 구글 생태계와 AI의 시너지 효과는 Gemini를 중심으로 1) 검색, 2) 플랫폼(Youtube), 3) 디바이스, 4) B2B 영역이 유기적으로 결합되는 양상을 보임
- 검색의 영역에서는 Gemini의 답변을 실시간 검색 데이터와 연동하여 환각 효과를 최소화하고 기업용 데이터의 신뢰도를 확보함
- 플랫폼에서는 Youtube와 멀티모달 AI인 Gemini가 통합되어 미디어 기업과 유튜버의 제작 자동화를 지원할 가능성이 높아짐
- 특히, 구글의 안드로이드 생태계는 약 39억 대의 기기에 Gemini를 기본 탑재하여 기존 구글 어시스턴트를 대체했으며, 앱 개발자에게 API 접근성을 높여 모바일 생태계를 장악할 수 있음
- 마지막으로 B2B 영역에서는 구글 클라우드와 워크스페이스(Workspace)는 통합되었으며, Google Docs, Sheets, Gmail 등의 업무 도구에 AI(Gemini)를 적용하므로써 기업용 AI 운영체제로서 영향력이 확대될 것으로 예상

Gemini 3.0이 적용된 구글 검색 AI 모드



자료: 대신증권 Research Center

구글 클라우드 플랫폼과 워크스페이스 관계

항목	Google Cloud Platform(GCP)	Google Workspace	통합
주요 역할	인프라, 서버, 데이터 분석, AI 개발	업무 생산성 및 문서 협업	엔터프라이즈 솔루션
형태	IaaS/PaaS	SaaS	계정 통합
핵심 연결 고리	Vertex AI, Big Query	Docs, Drive, Gmail 등	Gemini(AI 에이전트)

자료: 구글, 대신증권 Research Center

II. AI 기술 동향: 구글의 AI 검색 기술

① 구글의 강점은 AI 검색 기술(Retrieval Tech)

② AI 검색의 매커니즘, 쿼리 팬 아웃 기술

③ 일반 웹 검색 대비 서버 비용 증가 구조

④ 고효율 멀티 벡터 검색 기술 MuVERA로 검색 비용 효율화

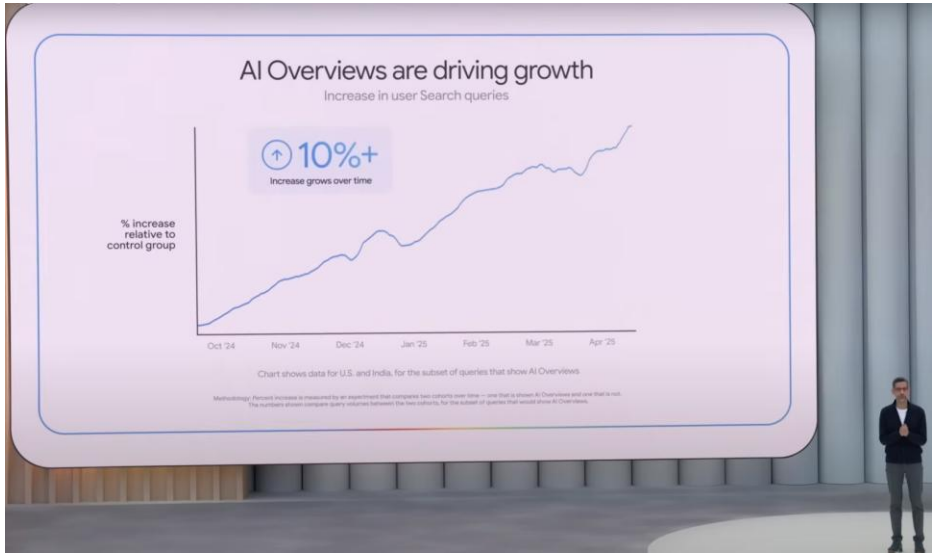
⑤ 구글의 강점은 AI 검색 기술(Retrieval Tech)

2. AI 기술동향

①구글의 강점은 AI 검색 기술(Retrieval Tech)

- 올해 5월 구글은 Google I/O에서 AI Overviews 덕분에 미국과 인도 같은 시장에서 검색 사용량이 10% 이상 증가했으며, AI 답변으로 인해 만족도가 높아져서 더 자주 검색하게 된다는 것을 데이터로 증명함
- AI 검색인 AI Overviews, AI 모드, Gemini에 구동되는 차세대 검색 인프라 기술인 쿼리 팬 아웃과 MuVERA가 도입된 것으로 추정
- 구글의 AI 검색은 사용자의 복잡한 질문을 다수의 하위 질의로 분해하는 쿼리 팬 아웃 기법을 적용함. 쿼리 팬 아웃 기술은 하나의 질문을 5~10개 하위 질문으로 확장하기 때문에 추론 및 검색 비용이 선형적으로 급증함
- 이와 같은 병목을 해소하기 위해 구글은 MuVERA 기술을 공개함. MuVERA는 다중쿼리·다중 벡터 구조에서 발생하는 연산 비용을 낮추면서 품질 저하 없이 검색 성능을 유지하는 것을 목표로 설계된 것으로 추정됨

AI Oveview로 인한 검색량 증가



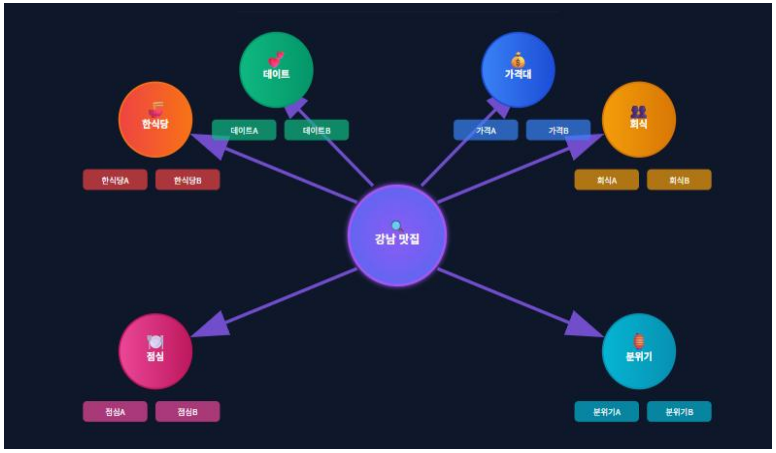
자료: 구글 I/O Keynote, 대신증권 Research Center

2. AI 기술동향

ⓐAI 검색의 매커니즘, 쿼리 팬 아웃 기술

- 쿼리 팬 아웃은 사용자의 단일 검색 쿼리를 여러 개의 관련 하위 쿼리로 분해하는 AI 기반 검색 기술이며, 모델의 아키텍처가 아닌 에이전트 방식으로 일하는 워크플로우 기술임
- 기존 검색은 사용자에게 웹사이트 및 정보를 제공해 사용자가 직접 필요한 정보를 찾는 방식이었지만, AI 검색의 쿼리 팬 아웃 기술은 사용자 질문의 의도를 먼저 탐색함. 그렇기 때문에 쿼리 팬 아웃 기술은 사용자의 한가지 질문에 다수의 쿼리를 병렬로 실행함
- 예시) 아이슬란드 여행 계획을 AI 검색으로 실행 시, 테마 1) 추천 여행 코스, 테마 2) 계절별 여행 팁, 테마 3) 예산 및 비용, 테마 4) 날씨 등 AI 검색 엔진이 자동으로 여러 테마를 생성해 다수의 쿼리를 병렬로 수행함

AI 검색의 쿼리 팬 아웃 작동 방식



자료: narr.enterprise, 대신증권 Research Center

기존 검색 방식과 쿼리 팬 아웃 기술 비교

항목	기존 검색 방식	Gemini 3.0 + 쿼리 팬 아웃
검색 수행 방식	한 개의 쿼리	다수 쿼리 병렬 실행
처리 목표	웹 사이트 및 정보 제공	질문 해결
결과 형태	링크 중심 및 사용자가 직접 필요한 정보 취사 선택	요약 + 시각화 + 액션
누락 가능성	높음	의도 기반 보완 탐색

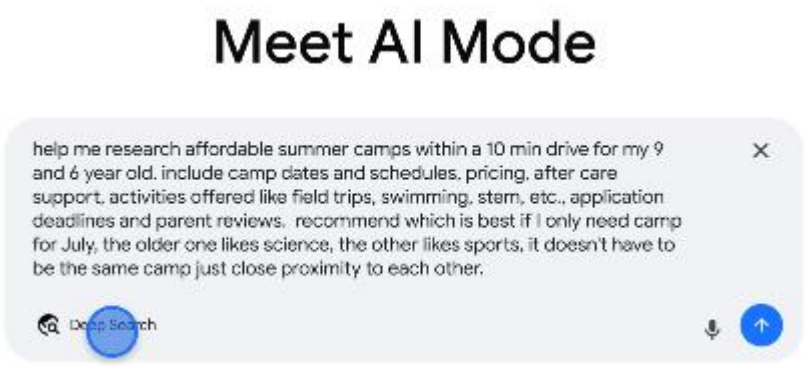
자료: 구글, 대신증권 Research Center

2. AI 기술동향

ⓒAI 검색의 매커니즘, 쿼리 팬 아웃 기술

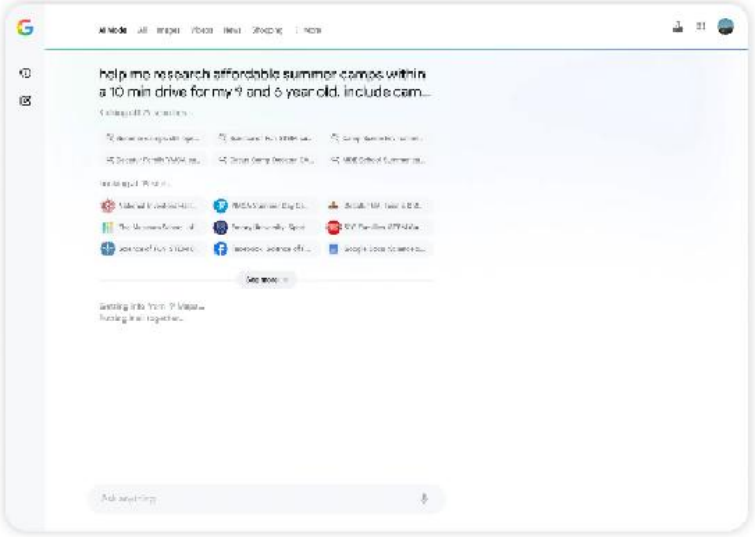
- 구글의 Query Fan-Out 특허는 하나의 질문을 여러 의도 기반 쿼리로 자동 확장하고, 검색 결과를 주제별로 요약·제시하는 새로운 AI 검색 인터페이스 기술임. 특히 검색된 문서들의 내용을 분석해 공통된 테마(Theme)를 자동으로 생성하고 분류해 제공함
- **사용자의 질문(1 Query) → 다수 쿼리 팬 아웃 → 테마 생성 → 클러스터링(Clustering) → 주제별 요약 제공**
- 테마 생성에는 LLM을 사용해 확보된 문서들의 내용을 분석해 비정형 데이터에서 의미있는 주제와 요약 설명을 생성함. 클러스터링의 경우 생성된 요약 설명들을 유사도 기반으로 그룹을 만들고 각 그룹의 중심에 가까운 표현을 그룹의 테마 명칭으로 선정함
- 구글 제미나이 딥 리서치 또한 쿼리 팬 아웃 기술을 활용 중이며, 결국 [질문 입력 → 여러 생각을 동시에 확장 → 압축된 결과 제공]이 현재 AI 검색 시대의 기본 패턴으로 자리 잡게 됨

구글의 AI 모드로 딥리서치



자료: 구글 제미나이, 대신증권 Research Center

구글 제미나이 딥 리서치 작동 모습, 다수의 쿼리가 실행됨



자료: 구글 제미나이, 대신증권 Research Center

2. AI 기술동향

③일반 웹 검색 대비 서버 비용 증가 구조

- 쿼리 팬 아웃 기술은 하나의 질문을 5~10개 하위 질문으로 확장하기 때문에 추론 및 검색 비용이 선형적으로 급증함
- 예시) AI 검색 질문 시, 1) Fan-Out: 검색 API 5회 호출 = $0.025, 2) 스크래핑: 상위 5개 총 25개 웹페이지를 가져옴 = $0.01, 3) 요약: 25개 페이지 텍스트(약 3만토큰)을 LLM에 입력 =LLM Input (GPT-5 급): (30k)*($2/1M) = $0.075, 4) 최종 답변 생성(500 토큰) = $0.005.
- 총 비용 질문 1회당 약 0.115달러(약 160원). 일반 챗봇(약 0.007~0.014달러) 대비 8~16배 , 일반 구글 검색 (약 0.003~0.006달러) 대비 10~30배 비싼 것으로 추정
- AI 검색의 비용 증가율은 일반 챗봇 대비 최소 8~16배 이상이지만 **사용자 만족도와 체류시간 증가로 보상 가능**
- 체류 시간이 증가할 수록 검색 점유율 확보를 통한 **광고 수익 증대** (주제 단위 추천 광고 가능), **상업적 타입의 검색 강화**(구매 전환 Funnel 최적화), **데이터 네트워크 효과**(광고 타게팅 성능 향상)이 가능함

AI 검색 단계 별 비용 증가 구조 예시

항목	핵심 비용 요소	기술적 부하 원인	추정 비용
쿼리 팬 아웃 (Query Fan-Out)	• 검색 API 호출 • **선형 증가: 1→N**	• 사용자 질문 1개를 6~10개의 하위 검색어로 분해 • 구글/Bing API는 호출 횟수 당 과금	• 구글 커스텀 검색 API: $5/1,000 리퀘스트 • Bing Search API: $15~25/1,000 트랜잭션
테마 랭킹 (Theme Ranking)	• 재정렬(Reranking) 연산 비용 • GPU Compute	• AI가 가져온 수십~수백 개 웹페이지의 연관성 평가 • 정확도가 높은 Cross-Encoder 모델 사용 → **연산량 수십 배 높음**	• Cohere Reranking API: $2/1M tokens • AWS P4d 인스턴스(Self-hosting): 시간당 $32
클러스터링 엔진 (Clustering Engine)	• 벡터화 & Vector DB • 메모리/ 스토리지	• 문서를 벡터화 하고 그룹화(클러스터링) 하는 과정 • 실시간 텍스트 벡터 변환 비용 + 벡터 DB 읽기/쓰기/ 비용 발생	• 오픈 AI Embedding: $0.02~0.13/1M tokens • Pinecone(Vector DB): 시간당 $0.07~0.10
요약 생성 (Summary Generation)	• LLM 입력 토큰 폭증 • Context Window 비용	• **가장 큰 비용 발생 구간** • 검색된 10~20개 문서의 본문을 LLM이 읽고 요약 • 출력보다 입력 토큰 양이 압도적으로 많음	• GPT-5: $2.00/1M input tokens (추정) • Gemini 3.0 Pro: $3.00~5.00/ 1M input tokens
맞춤형 인터페이스 (UI Personalization)	• Frontend/Edge 연산 • Latency(지연) 최적화	• 사용자별 맞춤 UI 렌더링을 위한 서버 사이드 렌더링(SSR) 또는 Edge Function 호출 비용	• Vercel/Cloudflare Workers: $0.15~0.50/1M 리퀘스트

주:AWS P4d 인스턴스는 클라우드에서 빌려 쓰는 고성능 AI 슈퍼컴퓨터 노드(초고성능 서버)
자료: Cohere, 오픈 AI, 구글, 대신증권 Research Center

2. AI 기술동향

④고효율 멀티 벡터 검색 기술 MuVERA로 검색 비용 효율화

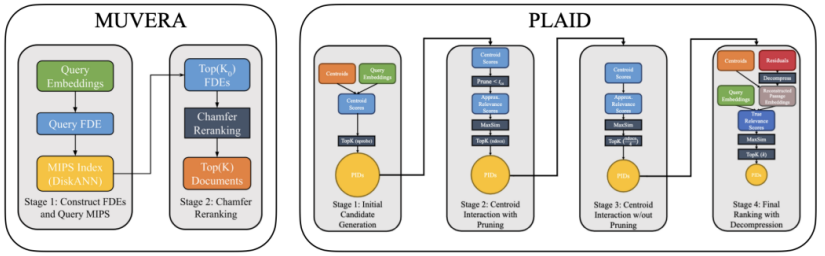
- 구글의 AI 검색은 기존 검색 대비 약 10~30배나 비싸기 때문에, AI 검색 비용을 최대한 낮춰 수익성 방어를 하려면 MuVERA와 같은 경량화 기술이 필요
- MuVERA(Multi-vector Retrieval via Fixed Dimensional Endocing) 기술은 보통 정확도는 높지만 느린 멀티 벡터 검색을 단일 벡터 검색만큼 빠르게 만드는 기술임. **올해 구글 I/O에서 발표한 MuVERA 검색 알고리즘은 검색 속도와 정확도를 동시에 향상시키는 기술임**
- 단일 벡터 검색이란 데이터를 하나의 벡터(숫자 묶음)로 변환해 검색하는 방식으로 속도는 빠르지만 복잡한 내용을 담지 못해 정확도가 낮음
- 멀티 벡터 검색은 데이터를 여러 개의 벡터로 검색하기 때문에 정확도가 높지만 계산이 복잡해 속도가 느리고 비용이 많이 듦
- 쿼리 팬 아웃이 검색 범위를 넓히는 역할을 한다면, Muvera는 넓어진 범위 내에서 필요한 정보를 빠르고 정확하게 찾아내는 역할임

단일 벡터 검색, 멀티 벡터 검색 , MuVERA 기술 비교

항목	단일 벡터 검색 방식 (Single Vector)	다중 벡터 검색 방식 (Multi-Vector)	MuVERA 기술
처리 속도 (쿼리당)	~10~20ms	~200~500ms	~40~60ms
의미 이해력	제한적 문맥 압축 과정에서 정보 손실	우수 모든 단어 간 관계 파악	우수 핵심 키워드 중심 맥락 파악
계산 비용 연산량(FLOPs)	문서당 1개 벡터 저장	문서당 128~256개 벡터 저장	벡터 양자화/압축 기술 적용
스토리지 효율	최상	메모리/스토리지 비용 50배 이상 폭증	기존 다중 벡터 대비 1/10 수준으로 절감
벡터 검색 GPU 비용 (1M 쿼리 당)	0.3~0.5달러	3~6달러	1~2달러

주: 벡터검색 비용은 NVIDIA v100 GPU 기준
자료: Google Research "MUVERA" (arXiv:2405.19504) 및 MS MARCO 벤치마크 결과를 기반으로 재구성, 대신증권 Research Center

단순화된 MuVERA 구조와 기존 검색 프로세스인 PLAID 비교



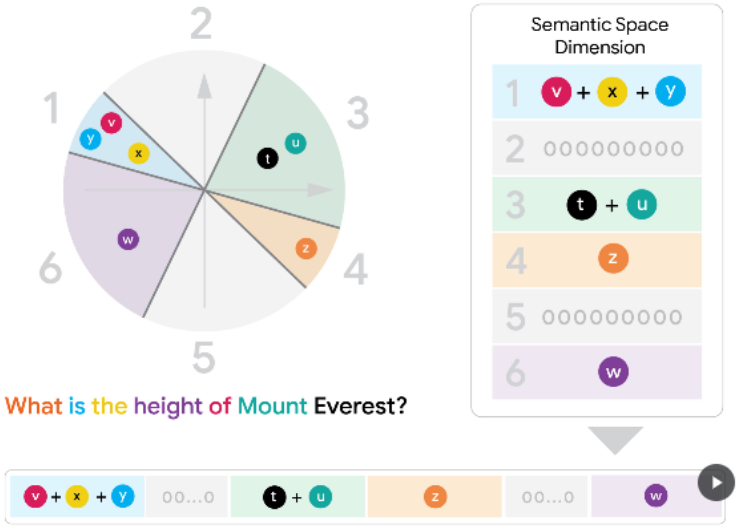
자료: 구글 리서치(arXiv:2405.19504), 대신증권 Research Center

2. AI 기술동향

④고효율 멀티 벡터 검색 기술 MuVERA로 검색 비용 효율화

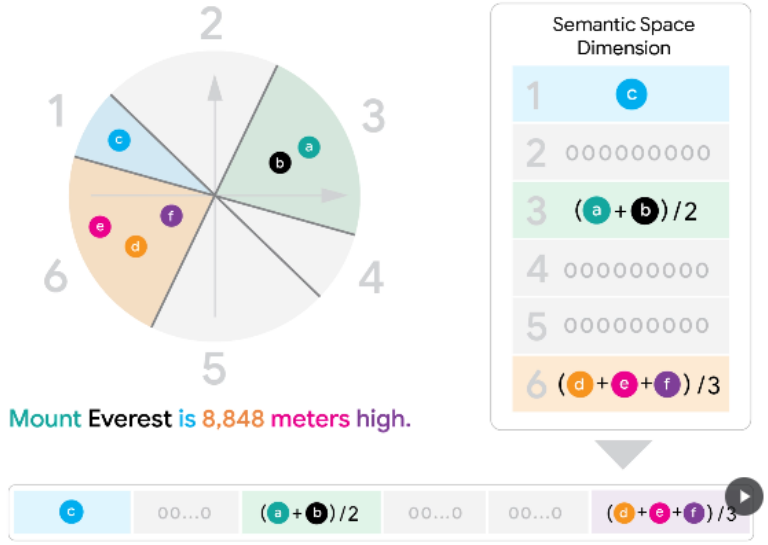
- 멀티 벡터 검색 기술인 MuVERA는 총 3단계로 1) 대략적 필터링(Coarse Filtering), 2) 핵심 정보 선별(Adaptive Pruning), 3) 지연 상호작용(Late Interaction)으로 구분
- 대략적 필터링 (Coarse Filtering)에서는 방대한 전체 문서에서 연관성 낮은 90% 데이터를 범위를 축소시킴
- 2단계인 핵심 정보 선별(Adaptive Pruning)에서는 질문의 핵심 의도(Intent)와 관련된 벡터만 남기고 불필요한 연산(토큰/벡터)을 제거하는 단계임
- 마지막인 지연 상호작용(Late Interaction)단계에서는 살아남은 상위 1% '핵심 벡터'와 '질문 벡터'를 세밀하게 1:1로 정밀 대조를 수행해 최종적으로 적합한 답변의 순위를 매김
- <u>결국 MuVERA 기술은 불필요한 연산 90%를 제거해 팬 아웃으로 늘어난 쿼리량을 상쇄하고 단순 키워드 매칭 대비 검색 정확도가 15% 이상 향상함</u>

사용자의 질문을 검색하기 쉬운 하나의 벡터로 요약하는 과정



자료: 구글리서치, 대신증권 Research Center

같은 방식으로 질문 또한 하나의 벡터로 압축됨



자료: 구글리서치, 대신증권 Research Center

III. AI 뉴스 & 데이터 대시보드

3. AI 데이터 대시보드

월간 AI 뉴스 정리

뉴스	내용
앤트로픽 CEO, 오픈AI 겨냥"데이터센터 투자 위험 지나치게 높아"	- 오픈AI의 초대형 인프라 프로젝트(1.4조달러), 코드 레드 발령 등 공격적 행보를 에둘러 언급하며 경제적 불확실성 속 투자 경쟁의 위험성을 강조 - 앤트로픽은 500억달러 규모의 데이터센터 투자를 진행 중이지만 상대적으로 보수적 접근을 강조
젠슨 황 "H200 수출 허용해도 중국이 받아들일지는 확신 못 해"	- 젠슨 황 CEO는 미국 정부가 H200 중국 수출 허용을 검토 중이지만, 중국은 성능이 낮춰진 칩을 받아들이지 않는다며 가능성을 낮게 평가 - 미국 의회가 '게인(GAIN) AI' 법안을 제외한 데는 환영 입장을 보였으나, H200 판매 허용은 미국 내 강경 반발이 거세다고 언급
조니 아이브와 샘 알트먼, 드디어 AI 하드웨어 프로토타입 완성했다고 밝혀	- 샘 알트먼과 조니 아이브가 OpenAI의 첫 AI 하드웨어 프로토타입을 개발 중이며, 화면 없는 스마트폰 크기의 단순하고 감성적인 디자인을 목표로 하고 있다고 밝힘 - 사용자가 직관적으로 끌리고 부담 없이 사용할 수 있는 형태를 강조하며, 제품이 2년 이내 출시될 수 있다고 언급
트럼프, 제2의 맨해튼 프로젝트 'Genesis Mission' 가동… AI로 과학 혁명 주도	- 미국 정부가 '제네시스 미션'을 출범해 연방 과학 데이터와 HPC를 통합한 AI 연구 플랫폼을 구축하고 과학 연구 자동화를 가속하려 함 - 안보·에너지·반도체 등 국가 과제를 AI로 해결하는 범정부 프로젝트로, 데이터·인프라 통합과 민관 협력, 강력한 보안 체계를 포함
오픈AI "챗GPT로 검색 10% 차지… 기능 강화 집중할 것"	- 챗GPT 총괄 닉 털리가 3주년 메시지에서 챗GPT가 전 세계 AI 비서 사용량 70%, 검색 활동의 10%를 차지한다고 강조하며 제미나이와의 경쟁 구도 속 우위를 부각 - 오픈AI 내부에서는 샘 알트먼의 '코드 레드'와 닉 털리의 '코드 오렌지' 발령으로 품질·속도·신뢰성·개인화 강화에 총력
구글, '제미나이 3' 출시로 시총 4조달러 눈앞	- 구글 주가가 5% 급등하며 사상 최고가(315.9달러)를 기록, 시총 3.82조달러로 4조달러에 근접 - 제미나이 3 초기 평가, 클라우드 성장 모멘텀, 버크셔 해서웨이 투자 참여 등이 주가 상승을 견인
구글, 'AI 모드'로 모바일 검색 전면 교체 시작	- 'AI 개요'와 'AI 모드'를 하나로 묶어 검색 결과 상단 요약 후 같은 화면에서 바로 대화형 후속 질문이 가능하도록 변경 - 사용자 흐름을 자연스럽게 대화 기반 탐색으로 전환해 제미나이 생태계 확장을 노림
젠슨 황 "AI 사용 줄이라는 것은 정신 나간 일…AI 때문에 인원 안 줄여"	- 젠슨 황 CEO는 내부 회의에서 일부 관리자가 직원에게 AI 사용을 줄이라고 한 데 강하게 비판 - AI 도입이 인력 감축으로 이어지지 않을 것이라며 지속적 채용 확대를 약속하고, 직원들이 AI 사용을 통해 기술 개선에 참여해야 한다고 강조
HSBC "오픈AI, 2030년까지 최소 300조 추가 펀딩 필요"	- HSBC는 오픈AI가 초대형 클라우드 임대 계약으로 인한 막대한 비용 때문에 2030년까지 최소 2070억달러가 부족할 것으로 분석함. - 사용자 증가·유료 전환 확대로 개선 여지는 있지만, 파트너사 리스크 등으로 장기 자금 조달 불확실성이 크다고 평가함
중국 반도체 거물 "14나노 중국산 AI 칩도 엔비디아 '블랙웰'에 근접 가능"	- 웨이샤오쥔 칭화대 교수가 3D 스태킹·하이브리드 본딩 기술을 활용하면 중국 기술로 블랙웰 GPU에 근접 가능하다고 발언 - 그러나 제시된 성능(와트당 2TFLOPS)은 블랙웰 대비 5배 낮아 실제 구현 가능성은 '개념적 모델' 수준이라는 평가

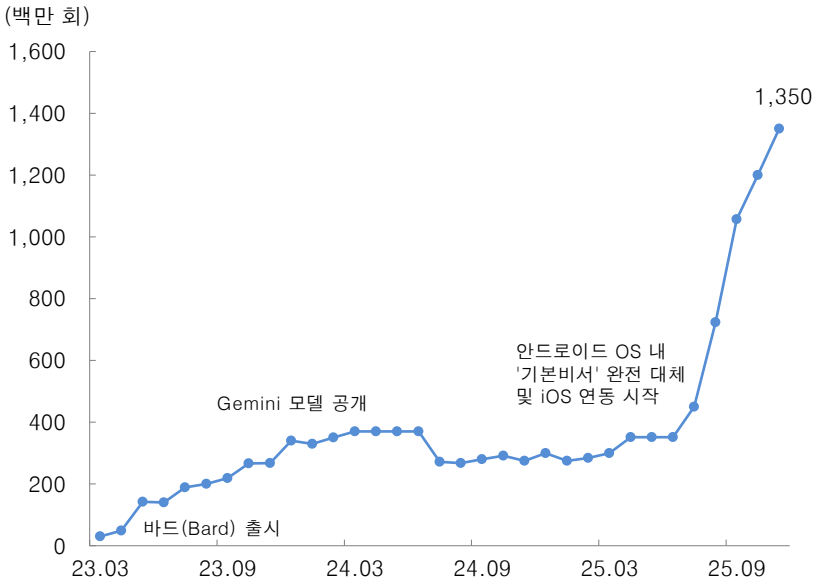
자료: 각 언론사 취합, 대신증권 Research Center

3. AI 데이터 대시보드

글로벌 생성형 AI 챗봇 점유율

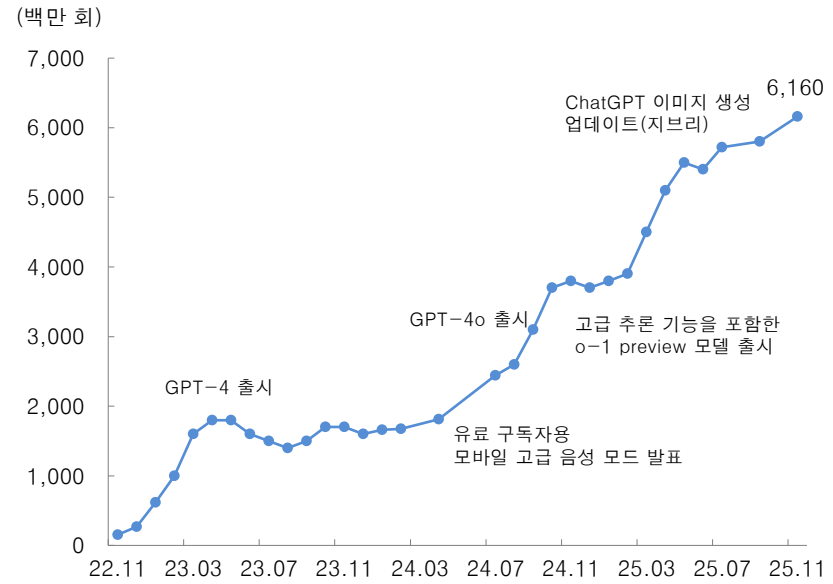
- 글로벌 생성형 AI 챗봇 중 ChatGPT는 약 81.8%의 점유율로 압도적인 1위를 기록 중. 그 뒤를 퍼플렉시티가 11.1%, 마이크로소프트(코파일럿)가 3.1%, 구글 제미나이가 약 3.0%로 뒤따르고 있음
- 2025년 11월 기준 ChatGPT 웹사이트의 월간 방문자 수는 약 62억 회을 기록했으며 제미나이의 경우 약 13억 회로 추정됨
- Gemini의 경우 올해 8월 안드로이드 OS내 '기본 비서'인 구글 어시스턴트를 Gemini로 전환 및 통합하면서 트래픽 급등하는 효과가 있었으며, 최근 11월 18일 Gemini 3.0이 공개되면서 트레픽이 상승할 것으로 예상됨

제미나이(Gemini) 월간 방문자 수 추이



주: 데스크탑과 모바일 합산
자료: Similarweb, Statista, 대신증권 Research Center

월간 ChatGPT 트래픽 추이



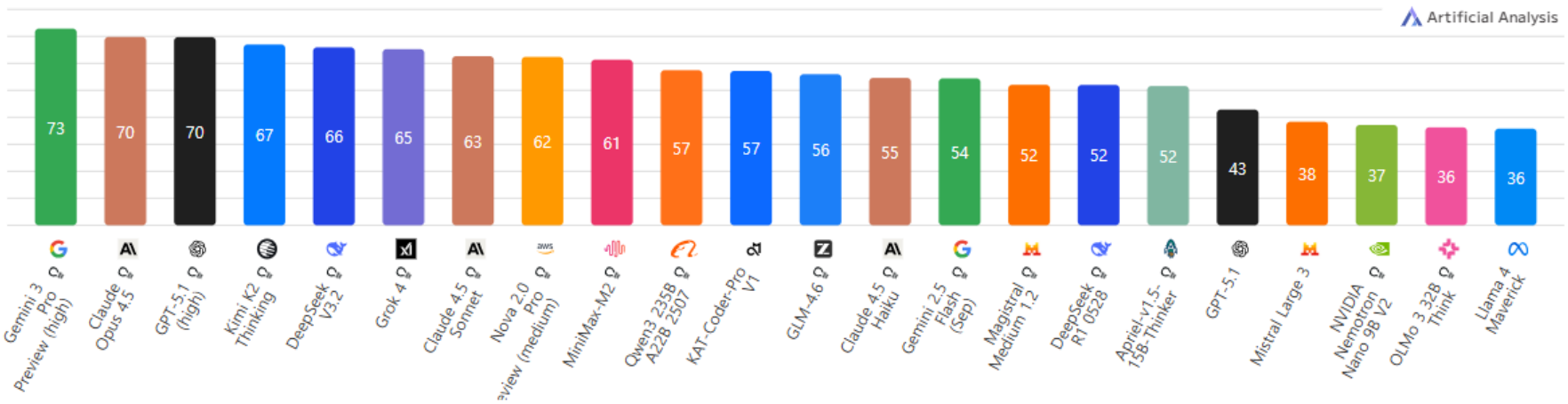
자료: Demandsage, 대신증권 Research Center

3. AI 데이터 대시보드

대형언어모델 성능 리더보드

- Artificial Analysis의 최신 LLM 리더보드에 따르면, 최근 출시한 Gemini 3.0 Pro가 73점으로 1위를 기록했으며, Claude Opus 4.5(70점), GPT-5.1(70점), Kimi K2 Thinking(67점), DeepSeek V3.2(66점)을 기록함
- Kimi K2 모델을 필두로 DeepSeek, Qwen 등 '중국 AI 3대장'이 GPT-5나 Gemini 3 등 미국 최상위 모델과 대등하거나 일부 지표에서 앞서는 성과를 보이고 있음
- 상위권 대부분은 여전히 폐쇄형(Closed-Source) 모델이 차지했으며, 미국과 중국은 여전히 초거대 LLM 경쟁에서 기술 성능 지표에서 선도적임을 보여주는 결과로 평가함

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

대형언어모델 성능 순위와 AI 모델의 속도 효율 순위

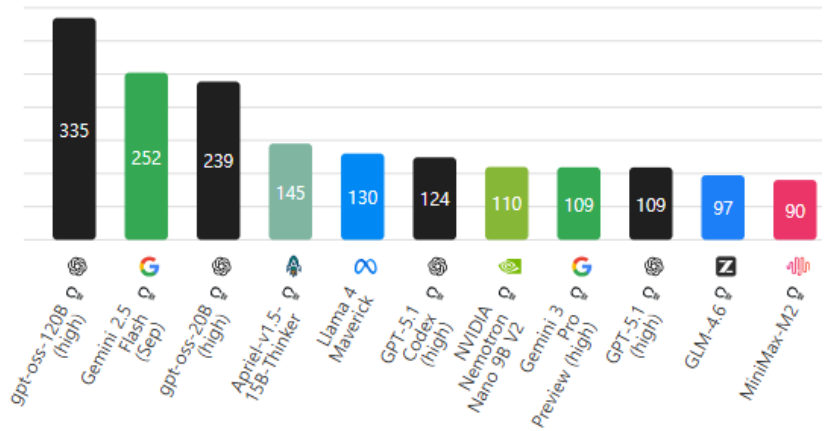
- LMArena의 최신 벤치마크에 따르면, 구글의 Gemini 3.0 Pro가 1,491점으로 전체 모델 중 가장 높은 점수를 기록함. 그 뒤로는 Grok 4.1과 Claude Opus 4.5 모델이 상위권을 차지했으며, 여전히 상위 10개 모델은 모두 폐쇄형(Proprietary) 모델로, 오픈소스 모델은 포함되지 않음
- 한편, 출력 속도(Output Speed) 부문에서는 OpenAI의 GPT-OSS-120B 모델이 초당 335 토큰(token/sec)으로 가장 빠른 성능을 보임. 뒤이어 Gemini 2.5 Flash가 252 token/s, Grok 4 Fast가 264 token/s를 기록함
- 특히 GPT-OSS-120B는 OpenAI가 공개한 부분 개방형(Open-Weight) 모델로, 단일 GPU에서도 구동 가능한 효율적 설계를 채택함. 이 모델은 추론, 코딩, 에이전트 작업 수행을 위한 경량형 구조를 갖췄으며, 연구 및 산업 응용에서도 높은 활용 가능성을 보임

대형언어모델 LMArena 벤치마크 성능 상위 10개 모델

순위	모델	점수	기업명	라이선스
1	gemini-3-pro	1,491	Google	폐쇄형모델
2	grok-4.1-thinking	1,481	xAI	폐쇄형모델
3	claude-opus-4-5-20251101-thinking-32k	1,467	Anthropic	폐쇄형모델
4	grok-4.1	1,465	xAI	폐쇄형모델
5	claude-opus-4-5-20251101	1,460	Anthropic	폐쇄형모델
6	gpt-5.1-high	1,457	OpenAI	폐쇄형모델
7	gemini-2.5-pro	1,452	Google	폐쇄형모델
8	claude-sonnet-4-5-20250929-thinking-32k	1,449	Anthropic	폐쇄형모델
9	claude-opus-4-1-20250805-thinking-16k	1,448	Anthropic	폐쇄형모델
10	claude-sonnet-4-5-20250929	1,445	Anthropic	폐쇄형모델

주: Overall 기준으로 수학, 지시 수행 능력, 다회 대화, 창작, 코딩, Hard Prompts 등의 기준을 모두 적용. 자료: LMArena, 대신증권 Research Center

AI 모델의 답변 속도 순위



자료: Artificial Analysis, 대신증권 Research Center

3. AI 데이터 대시보드

글로벌 AI 소프트웨어 관련 밸류체인

빅테크 & 클라우드	연도	ALPHABET	META PLATFORMS	ALIBABA GROUP	BAIDU	NAVER CORP	MICROSOFT	AMAZON	ORACLE	SOFTBANK
코드		GOOG US	META US	BABA US	BIDU US	035420 KS	MSFT US	AMZN US	ORCL US	9984 JT
거래소		NASDAQ GS	NASDAQ GS	New York	NASDAQ GS	Korea SE	NASDAQ GS	NASDAQ GS	New York	Tokyo
주가 (현지통화)		318	662	157	119	248,500	481	229	214	18,700
시가총액(십억달러)		3,837	1,668	376	42	26	3,574	2,449	611	172
주가상승률 (%)	'23	58.8	194.1	-12.0	4.1	26.2	56.8	80.9	29.0	4.1
	'24	35.1	65.4	9.4	-29.2	-11.2	12.1	44.4	58.1	30.9
매출액	24A	350,018	164,501	131,313	18,490	7,875	245,122	637,959	52,961	46,794
(백만달러)	25F	340,485	199,345	147,113	18,319	8,222	326,332	714,539	67,004	49,032
	26F	389,400	235,435	163,291	19,251	9,154	375,113	795,333	83,617	52,409
영업이익	24A	112,390	69,380	15,815	2,954	1,452	109,433	68,593	15,353	5
(백만달러)	25F	131,949	82,331	13,864	1,880	1,519	150,525	79,864	28,439	24,271
	26F	155,474	86,237	21,061	2,301	1,726	173,574	100,461	33,689	13,960
영업이익률	24A	32.1	42.2	12.0	16.0	18.4	44.6	10.8	29.0	0.0
(%)	25F	38.8	41.3	9.4	10.3	18.5	46.1	11.2	42.4	49.5
	26F	39.9	36.6	12.9	12.0	18.9	46.3	12.6	40.3	26.6
순이익	24A	100,118	62,360	11,163	3,300	1,410	88,136	59,248	10,467	-1,577
(백만달러)	25F	129,292	60,652	15,921	1,943	1,435	119,407	76,747	14,871	9,162
	26F	134,093	76,252	19,723	2,433	1,529	140,509	86,277	18,730	4,702
순이익률	24A	28.6	37.9	8.5	17.8	17.9	36.0	9.3	19.8	-3.4
(%)	25F	38.0	30.4	10.8	10.6	17.5	36.6	10.7	22.2	18.7
	26F	34.4	32.4	12.1	12.6	16.7	37.5	10.8	22.4	9.0
ROE	24A	32.9	37.1	8.1	9.1	7.9	37.1	24.3	214.1	-2.6
(%)	25F	33.2	29.7	10.2	5.2	7.8	29.4	22.1	55.7	19.0
	26F	31.2	28.9	11.5	6.0	7.7	27.3	19.8	41.4	7.6
PER	24A	-	24.4	10.2	10.1	15.7	37.8	39.0	30.2	-
(배)	25F	29.1	22.9	25.4	15.6	18.6	30.1	26.5	31.4	13.7
	26F	26.9	19.7	17.1	14.6	17.4	25.6	24.3	27.1	41.0
EPS	24A	-	24.61	4.41	9.23	9.32	11.86	5.66	3.82	-1.18
(달러)	25F	10.94	28.94	6.21	7.62	9.08	15.96	8.65	6.82	8.79
	26F	11.83	33.51	9.23	8.15	9.69	18.76	9.43	7.91	2.94
EV/EBITDA	24A	-	20.1	4.6	3.7	13.5	27.9	23.4	20.1	87.0
	25F	22.0	16.1	17.4	19.7	13.0	17.8	15.0	19.5	29.3
	26F	18.4	13.9	10.6	7.2	10.4	15.8	12.4	15.6	14.3

자료: Bloomberg, 대신증권 Research Center

3. AI 데이터 대시보드

글로벌 AI 소프트웨어 관련 밸류체인

AI Software	연도	SAP SE	PALANTIR	SALES FORCE	IBM	Service Now	INTUIT	PALO ALTO	TEMPUS AI	UIPATH	C3.AI	INNODATA	ADOBE	CROWD STRIKE
코드		SAP gr	PLTR US	CRM US	IBM US	NOW US	INTU US	PANW US	TEM US	PATH US	AI US	INOD US	ADBE US	CRWD US
거래소		Xetra	NASDAQ GS	New York	New York	New York	NASDAQ GS	NASDAQ GS	NASDAQ GS	New York	New York	NASDAQ GM	NASDAQ GS	NASDAQ GS
주가 (현지통화)		212	178	247	308	839	663	196	76	18	15	58	329	513
시가총액(십억달러)		303	424	232	288	174	185	136	14	10	2	2	138	129
주가상승률 (%)	'23	49.8	167.4	98.5	16.1	82.0	60.6	111.3	-	95.4	156.6	174.5	77.3	142.5
	'24	59.3	340.5	27.1	34.4	50.1	0.6	23.4	-	-48.8	19.9	385.5	-25.5	34.0
매출액	24A	36,971	2,866	34,857	62,753	10,984	16,285	8,028	693	1,308	311	170	21,505	3,056
(백만달러)	25F	43,157	4,394	41,437	67,059	13,242	21,203	10,528	1,265	1,583	300	249	23,684	4,802
	26F	47,746	6,279	45,865	70,174	15,696	23,838	11,924	1,575	1,727	332	311	25,899	5,861
영업이익	24A	8,821	310	5,011	8,384	1,364	3,630	684	-691	-165	-318	24	6,741	-2
(백만달러)	25F	11,947	2,136	14,122	13,121	4,112	8,660	3,130	-234	350	-198	41	10,935	1,037
	26F	13,936	3,019	15,802	14,068	4,944	9,761	3,614	-130	392	-165	49	11,863	1,414
영업이익률	24A	23.9	10.8	14.4	13.4	12.4	22.3	8.5	-99.7	-12.6	-102.5	14.3	31.3	-0.1
(%)	25F	27.7	48.6	34.1	19.6	31.1	40.8	29.7	-18.5	22.1	-66.1	16.3	46.2	21.6
	26F	29.2	48.1	34.5	20.0	31.5	40.9	30.3	-8.3	22.7	-49.6	15.6	45.8	24.1
순이익	24A	6,022	462	4,136	6,023	1,425	2,963	2,578	-706	-90	-280	29	5,560	89
(백만달러)	25F	8,207	1,461	7,026	8,506	1,855	4,446	1,533	-259	60	-436	31	7,070	-207
	26F	9,129	2,009	8,041	9,639	2,531	5,211	1,831	-234	74	-411	38	7,690	71
순이익률	24A	16.3	16.1	11.9	9.6	13.0	18.2	32.1	-101.8	-6.9	-90.1	16.8	25.9	2.9
(%)	25F	19.0	33.2	17.0	12.7	14.0	21.0	14.6	-20.4	3.8	-145.6	12.4	29.9	-4.3
	26F	19.1	32.0	17.5	13.7	16.1	21.9	15.4	-14.8	4.3	-123.7	12.1	29.7	1.2
ROE	24A	12.6	10.9	7.0	24.2	16.5	16.6	74.5	-	-4.6	-31.0	64.3	36.3	4.7
(%)	25F	15.5	28.2	15.0	36.3	26.1	28.3	24.8	-43.4	18.5	-21.5	-	61.0	20.5
	26F	17.1	29.4	15.2	31.9	24.9	28.0	21.8	-5.6	18.0	-21.5	-	63.4	18.4
PER	24A	88.2	385.4	53.5	19.4	150.9	58.9	122.8	-	-	-	44.4	36.3	719.3
(배)	25F	34.7	246.1	21.0	27.1	48.3	28.6	50.8	-	27.5	-	66.7	15.8	138.4
	26F	29.5	179.2	19.0	25.2	41.0	25.2	45.1	-	24.7	-	56.1	14.1	107.0
EPS	24A	5.17	0.21	4.25	6.53	6.92	10.58	4.04	-6.23	-0.16	-2.34	0.98	12.43	0.37
(달러)	25F	7.12	0.72	11.77	11.36	17.38	23.16	3.85	-0.62	0.67	-1.19	0.88	20.83	3.71
	26F	8.36	0.99	13.00	12.20	20.45	26.33	4.33	-0.47	0.75	-0.93	1.04	23.38	4.79
EV/EBITDA	24A	31.9	479.4	27.5	15.4	124.0	40.0	90.8	-	-	-	48.8	27.4	1,702.4
	25F	22.5	195.8	12.9	18.1	36.5	20.6	36.7	25,835.3	31.5	-	-	11.7	97.8
	26F	19.6	132.0	12.3	17.7	30.8	18.7	32.0	169.3	20.6	-	26.9	11.0	73.6

자료: Bloomberg, 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	10,565	0.80	2023.05.16	1.6	-6.6	14.3	33.3	41.9	30.0
A466950	TIGER 글로벌AI액티브	Indxx Artificial Intelligence and Big Data 지수(PR)	3,749	0.79	2023.10.11	1.7	-4.4	19.2	38.3	28.8	25.8
A481180	SOL 미국AI소프트웨어	KEDI 미국 AI 소프트웨어지수(PR)	2,403	0.45	2024.05.14	5.0	-4.4	7.9	11.8	14.5	14.1
A0040Y0	SOL 팔란티어커버드콜OTM 채권혼합	KEDI 팔란티어커버드콜OTM 채권혼합지수(NTR)	2,506	0.35	2025.04.22	1.2	-0.4	5.5	8.6	-	16.2
A0102A0	TIGER 미국AI소프트웨어TOP4Plus	Akros U.S. AI Software TOP4 Plus 지수 (Price Return)	1,330	0.49	2025.09.09	5.1	-8.4	-	-	-	6.3
A412770	TIGER 글로벌AI플랫폼액티브	Indxx Global Metaverse Index (Price Return)	1,260	0.79	2021.12.22	1.1	-5.5	7.0	15.9	20.5	15.9
A0040X0	SOL 팔란티어미국채커버드콜혼합	KEDI 팔란티어미국채커버드콜혼합지수(NTR)	1,067	0.35	2025.04.22	1.2	-1.3	12.1	22.1	-	33.6
A0041D0	KODEX 미국AI소프트웨어TOP10	iSelect 미국 AI 소프트웨어 TOP10 지수(PR)	756	0.45	2025.04.22	4.9	-4.7	12.4	23.8	-	61.4
A0093D0	KoAct 팔란티어밸류체인 액티브	KEDI 팔란티어밸류체인지수(PR)	454	0.50	2025.08.26	0.8	-4.9	24.6	-	-	24.5
A0047R0	RISE 팔란티어고정테크100	KEDI 팔란티어고정테크100지수(PR)	402	0.20	2025.05.13	3.1	-1.1	16.6	32.5	-	36.8
A0067Y0	TIGER 차이나AI소프트웨어	Mirae Asset China AI Software 지수 (Price Return, HKD)	265	0.49	2025.06.17	-2.0	-4.6	10.6	-	-	25.1
A0015F0	KIWOOM 팔란티어미국30년국채혼합액티브(H)	Bloomberg Blended PALANTIR and US Long Treasury Bond Index	243	0.19	2025.02.11	0.0	-3.4	5.1	10.5	-	11.4
A0105D0	SOL 한국AI소프트웨어	KEDI 한국AI소프트웨어 지수(PR)	209	0.45	2025.09.23	4.8	-4.2	-	-	-	3.3
A0050E0	PLUS 미국AI에이전트	Solactive US AI Agents Index(PR)	56	0.45	2025.05.20	4.9	0.3	8.0	3.3	-	1.5

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: 박세라)
- 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.