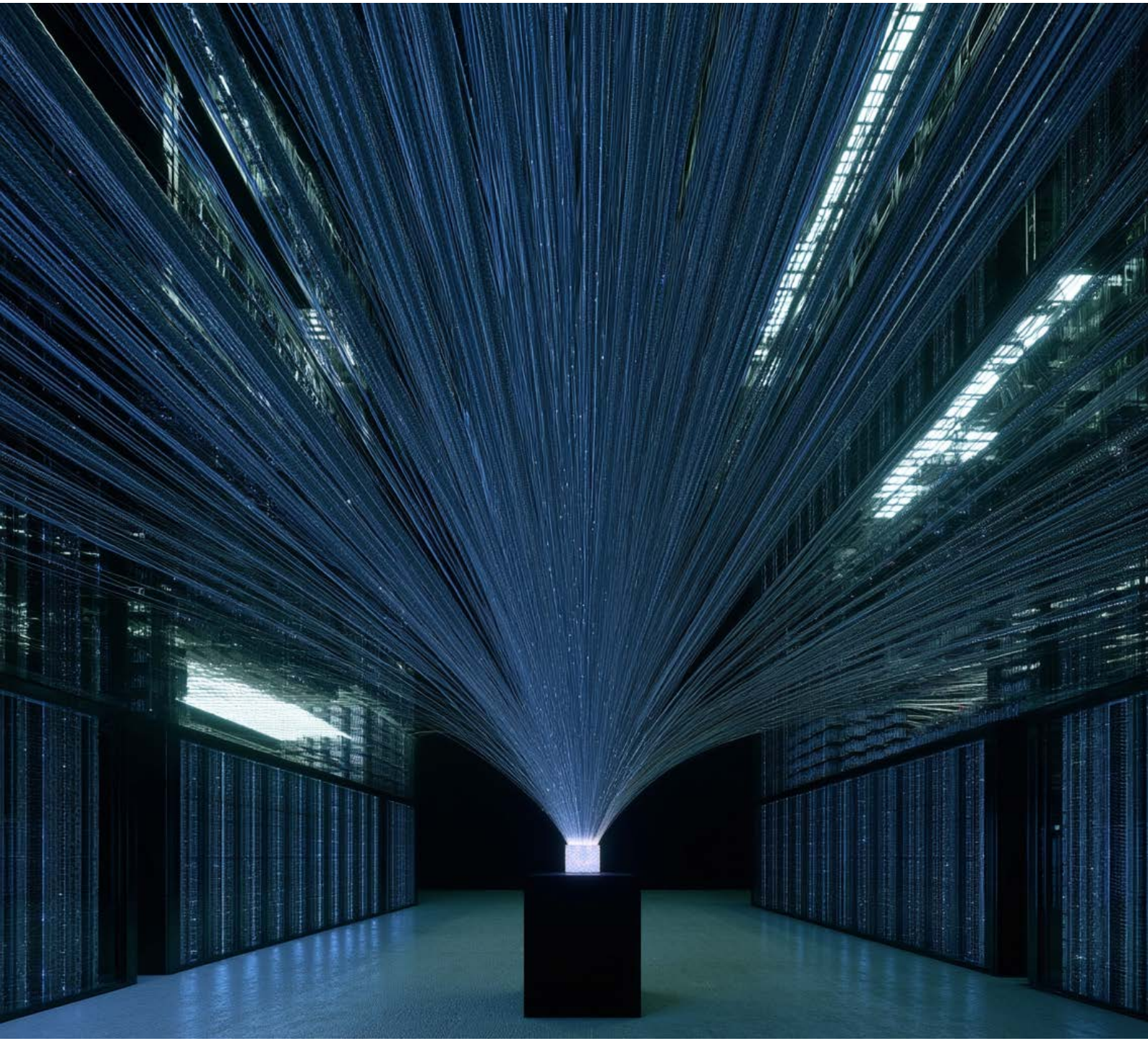
Market Insight | 2026.3.27

그림: Midjourney(무지막대한 데이터 압축이 낳은 지능의 폭발)

구글 TurboQuant 쇼크

제본스는 멘탈이 완전히 나가버렸습니다

한종목
chongmok.han@miraeasset.com



Executive Summary

메모리 슈퍼사이클의 끝이 아니다. 추론 인프라가 더 가속할 것이다.

TurboQuant의 등장은 시장에 다시 한번 익숙한 공포를 불러왔습니다. KV 캐시를 16비트에서 3~4비트 수준까지 압축할 수 있다면, 결국 AI가 필요로 하는 메모리 총량도 줄어들 것이고, 그렇다면 HBM과 서버 DRAM, 기업용 SSD를 둘러싼 낙관도 꺾여야 한다는 해석입니다. 실제로 이런 종류의 알고리즘 발표는 메모리 반도체 주가의 단기 센티멘트를 흔들기에 충분합니다. 그러나 **논문의 원본은 약 1년 전인 2025년 4월에 이미 공개되었고, 핵심 기초 연구자는 그보다 2년 전에 구글을 떠났습니다. 시장이 "뉴스"라고 부른 것은, 뉴스가 아니었습니다.**

TurboQuant는 수학적으로 우아한 압축 기법입니다. 무작위 회전으로 데이터 분포를 균일화하고(PolarQuant), 극좌표 변환으로 보정 상수를 제거한 뒤, 1비트 잔차 처리(QJL)로 편향을 상쇄합니다. 3.5비트에서 16비트 원본과 정확히 동일한 성능(LongBench 기준)을 달성했으며, H100 GPU에서 최대 8배의 속도 향상을 이뤄냈습니다. 엣지 디바이스에서의 배포 장벽을 허문 것은 사실입니다.

그러나 프로덕션 환경은 이미 FP8/FP4로 돌아가고 있어, 실**질적 추가 이득은 6배가 아니라 약 2.7배에 불과합니다. DeepSeek MLA의 28배 압축이 이미 존재하고, 추론 시 양자화는 업계에서 "이미 해결된 문제"로 분류됩니다. TurboQuant는 점진적 개선이지, "제로-투-원"이 아닙니다.**

효율화가 수요를 줄인 적이 있습니까? 단 한 번도 없습니다. 2026년 3월 바로 지금, 제본스 역설의 가장 생생한 증거가 동시다발적으로 벌어지고 있습니다.

Anthropic의 Claude는 주간 30% 성장하며 세션 리밋을 조이고 있고, OpenAI는 Sora에 태우던 연 $54억의 GPU를 Codex로 재배치하면서 AI 비디오 생성 시장을 xAI의 Grok에게 내줬습니다. 효율화 기술이 이미 존재하는데도 컴퓨팅 수요가 공급을 초과하고 있다는, 2026년 3월 시점의 가장 직접적인 증거입니다. 만약 TurboQuant가 정말로 비용을 낮추면, 샘 알트만은 Sora를 다시 부활시킬 것입니다. 이는 **지금 이 순간에도 컴퓨팅이 남아도는 것이 아니라 여전히 절대적으로 부족하다는 현실을 보여줍니다. 효율화는 축소가 아니라, 과거에 비용 때문에 포기했던 워크로드를 다시 부활시키는 연료입니다**.

결론적으로 이번 TurboQuant 이슈는 메모리 슈퍼사이클의 종말 신호라기보다, 추론 인프라가 다음 단계로 넘어가기 위한 비용 구조 개선의 신호로 해석하는 편이 타당합니다. 중장기적으로 보면, AI 인프라의 물리적 토대를 공급하는 HBM과 고용량 DRAM, 기업용 스토리지의 전략적 가치는 오히려 더 선명해질 가능성이 높습니다.

10년간 AI 추론 인프라를 구축하며 모든 양자화를 써본 AI 베테랑 개발자들의 중론은, "메모리는 여전히, 그리고 앞으로도 병목입니다." 제본스의 역설은 산업의 당사자가 인정하고, 주가가 증명했었던 구조적 현실입니다.

I. TurboQuant: 제본스의 역설, 그 두 번째 증명

1. TurboQuant는요, 사실 작년 봄에 나온 거예요

(1) 센티멘트는 언제나 펀더멘털보다 먼저 흔들린다

시장에는 늘 이런 순간이 있다. 기술 발표 하나가 나오면, 사람들은 그 기술이 현실의 어느 층을 건드리는지 보기 전에 먼저 극단적인 결론부터 가져다 붙인다. 이번에 구글이 내놓은 TurboQuant도 비슷하다.

그림 1. 2026년 3월 24일에 게시된 구글 리서치의 공식 블로그 포스트에 올라온 TurboQuant '극단적 압축으로 AI 효율을 재정의한다'는 자극적인 헤드라인을 달고 있음.
그런데, 이 발표의 기저에 깔린 알고리즘은 이미 1년 전인 작년 4월에 논문으로 세상에 던져진 기술

자료: 구글, 미래에셋증권 리서치센터

KV 캐시를 기존의 16비트에서 3~4비트 수준까지 압축할 수 있다는 문장이 나오자마자, 일부 시장은 곧장 "메모리를 덜 쓰게 된다 → 메모리 반도체 업황은 꺾인다"는 논리에 매달렸다. 이러한 반응은 직관적으로는 쉬워 보이지만, 실제 AI 시스템이 어디에서 돈을 태우고 있는지, 그리고 절감된 비용이 어디로 다시 흘러가는지를 생각하면 좀 얕다고 생각된다.

문제는, 그 해석이 지나치게 평면적이라는 데 있다. AI 인프라의 병목은 생각보다 훨씬 구조적이다. 2026년 현재 하이퍼스케일러들은 AI 인프라 구축에 막대한 자본을 태우고 있고, 그 지출의 중심은 더 이상 학습만이 아니다. 추론이 전체 컴퓨팅 지출에서 차지하는 비중이 급격히 높아지면서, 이제 핵심 병목은 "얼마나 많이 계산하느냐"보다 "얼마나 많은 상태를 메모리에 유지하고 다시 불러오느냐"로 이동했다. 그리고 그 중심에 KV 캐시가 있다.

그래서 이 글의 출발점은 수학적 우아함 자체가 아니다. 더 중요한 것은, TurboQuant라는 기술이 메모리 반도체 업종에 어떤 종류의 오해를 만들 수 있는가다. 주식 시장은 기술을 정확히 이해한 뒤 움직이지 않는다. 먼저 움직이고, 나중에 논리를 붙인다. 그렇기 때문에 "TurboQuant가 메모리 수요를 죽인다"는 식의 단순 서사는 매우 빠르게 퍼질 수 있다.

하지만 이 사안을 그렇게 읽으면, 정작 더 중요한 흐름이 가려진다. TurboQuant는 메모리 수요를 철거하는 기술이 아니라, 추론 병목의 가장 비싼 조각 하나를 다시 배선하는 기술이다. 다시 말해, AI 인프라 전체를 축소하는 기술이 아니라 추론 인프라를 더 공격적으로 확장할 수 있게 만드는 기술에 가깝다. 시장은 지금 이 둘을 혼동하고 있다.

2. TurboQuant의 실체: 메모리 전체가 아니라 추론의 KV 캐시를 겨눈다

(1) TurboQuant가 줄이는 것은 '모든 메모리'가 아니다

핵심부터 말하면, TurboQuant는 AI 시스템 전체의 메모리를 줄이는 기술이 아니다. 정확히는 추론 과정에서 대화가 길어질수록 급격히 커지는 KV 캐시를 더 작게 저장하기 위한 기술이다. 여기서 출발점을 잘못 잡으면, 뒤의 해석도 전부 엇나간다.

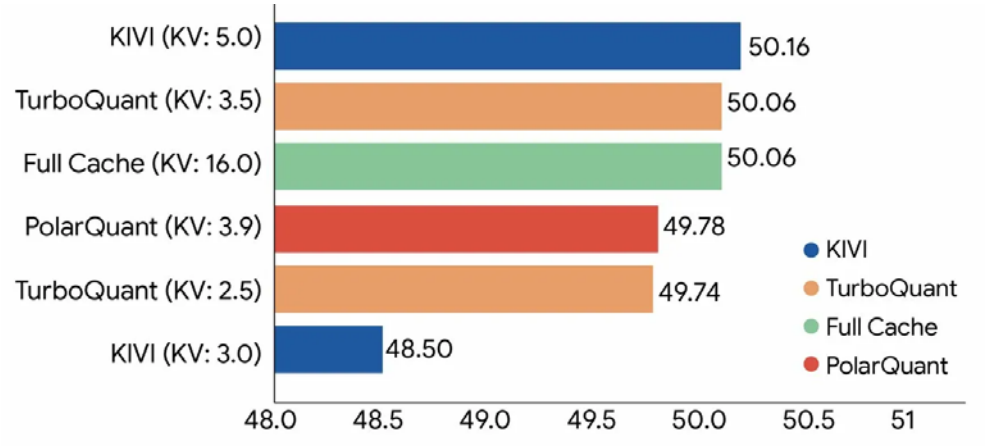
트랜스포머 모델은 사용자의 입력을 처리할 때, 이전 토큰들의 Key와 Value를 계속 저장해 둔다. 이것이 KV 캐시다. 문제는 문맥이 길어지고 사용자 수가 늘어나고 에이전트가 동시에 여러 작업을 돌릴수록, 이 캐시가 매우 빠른 속도로 불어난다는 점이다. 결국 추론 비용에서 중요한 병목 중 하나는 연산량 자체보다, 이 KV 캐시를 얼마나 효율적으로 저장하고 다시 읽어오느냐에 있다. TurboQuant는 바로 그 지점을 겨눈다.

(2) 핵심은 데이터를 "더 세게"가 아니라 "더 잘" 압축하는 방식이다

기존에도 KV 캐시를 줄이려는 시도는 많았다. 하지만 너무 세게 압축하면, 나중에 다시 꺼내 쓸 때 정보가 많이 깨져서 모델의 답변 품질이 급격히 떨어졌다. 문제는 단순히 비트를 줄이는 것이 아니었다. 작게 만들면서도, 어텐션 계산에 필요한 의미 정보와 상대적 거리 구조를 최대한 보존해야 했다. 이게 어려웠다.

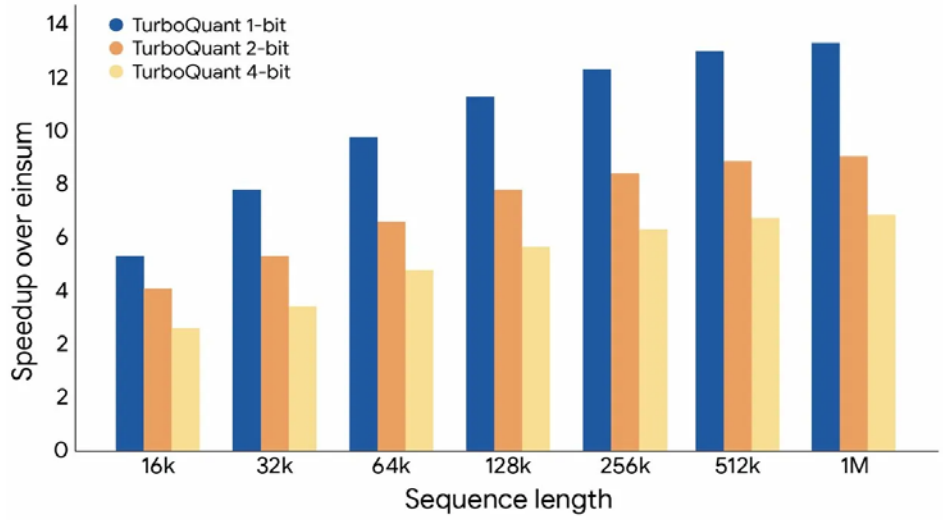
TurboQuant가 의미 있는 이유는 바로 이 지점을 개선했기 때문이다. 이 기술은 데이터를 무작정 눌러 담는 것이 아니라, 먼저 압축하기 쉬운 형태로 다시 정리한 뒤 더 작은 단위로 저장하고, 그 과정에서 생기는 오차도 따로 보정한다. TurboQuant의 핵심 기술 중 하나인 PolarQuant는 원래 불규칙한 고차원 벡터를 "무작위 회전과 극좌표 변환"을 통해 보다 다루기 쉬운 분포로 정렬한다. 이로써 기존 양자화가 안고 있던 스케일링 메타데이터 오버헤드를 크게 줄인다. 또 다른 핵심인 QJL은 그렇게 압축한 뒤 남게 되는 잔차 정보를 아주 작은 차원 수준으로 보정해, 어텐션 점수에서 생기는 체계적 편향을 억제한다.

그림 2. KIVI / TurboQuant / Full Cache 성능 비교 바 차트
KIVI 5bit보다도 더 낮은 비트수에서 비슷하거나 더 나은 성능을 낸다는 점이 핵심
사람들이 놀란 이유는 압축만 했기 때문이 아니라, 이 정도 압축에서도 성능이 안 무너졌기 때문



자료: 구글, 미래에셋증권 리서치센터

그림 3. Sequence length별 TurboQuant speedup(기존 방식보다 몇 배 빨라졌는가) 차트
16k보다 뒤쪽인 1M 쪽으로 갈수록 커지는 것은, 이 기술이 긴 문맥의 병목을 겨누고 있음을 의미
다시 말해, 짧은 프롬프트가 아니라 긴 세션과 에이전트 워크플로우에서 더 강해지는 기술이라는 것
TurboQuant는 지금까지 비싸서 돌리기 힘들던 초장문 추론을 열어주는 기술일 수 있음



자료: 구글, 미래에셋증권 리서치센터

무슨 말인지 당연히 이해하기 어렵다. 3D 그래픽스 업계의 기술자들이 쓰던 생소한 수학 개념이기 때문이다. 실제로도 그래픽 엔지니어는 "이걸 이제야 발견했나"라고 말하기도 한다. 어쨌든 이 과정을 좀 쉽게 쓰자면 다음과 같다.

기존 방식은 데이터를 아주 작은 블록으로 쪼개 억지로 구겨 넣는 방식으로 압축을 한다. 그리고 나중에 압축을 풀어보기 위한 꼬리표를 매번 같이 들고 다녀야 했다. 각 블록마다 최소값과 최대값을 찾아 "이 블록의 숫자는 이 범위 안에 있다"는 꼬리표다. 문제는 그 꼬리표 자체가 메모리를 먹는다는 점이다. 데이터를 줄이려고 했는데, 데이터를 어떻게 줄였는지 설명하는 메타데이터 때문에 또 공간을 쓴다. 양자화의 저주가 바로 여기서 시작된다.

그래서 구글이 내놓은 압축 발상은 개별 블록마다 다른 눈금자를 쓰지 말고, 아예 모든 데이터를 비슷한 분포처럼 보이게 만든 뒤, 공통의 눈금자 하나로 다뤄 버리자는 것이다. 이렇게 되면 블록마다 최소/최대값을 따로 기록할 필요가 줄어든다. 더 쉽게 말하면, 각 데이터 상자가 어디에 있을지를 가늠하기 위해 각각 다른 자를 대는 방식에서 벗어나, 모두가 공통으로 된 하나의 자로 바꾸는 것이다. 그러니 이런 저런 데이터 병목이 사라진다.

TurboQuant는 그 꼬리표 비용을 줄이고, 압축 과정에서 생기는 왜곡도 최소한의 추가 정보만으로 보정한다. 그래서 16비트에서 3~4비트까지 압축해도 성능 붕괴 없이 버티는 것이다.

즉, 기존에는 실전에서 쓰기 어려웠던 초강한 압축을 실제 추론 환경으로 끌고 들어왔다는 점에 주안점을 둬야 한다. 혁명은 아닐 수 있어도 충분히 산업적이다. 정확히 작년 DeepSeek 쇼크 때가 생각이 난다.

(3) "6배 압축"이라는 헤드라인을 그대로 믿지는 말아야 한다

여기서 시장이 자주 놓치는 함정이 있다. 벤치마크 숫자가 곧바로 산업 전체의 경제성으로 번역된다고 믿는 것이다. 구글 논문의 모든 비교는 '16비트 풀 캐시(FP16)'를 기준선(Baseline)으로 삼고 있다는 것이다.

문제는 2026년 현재, 프로덕션 환경에서 아무도 KV 캐시를 16비트로 저장하지 않는다는 사실이다. vLLM, SGLang, TensorRT-LLM 같은 주류 서빙 프레임워크는 이미 FP8(8비트) KV 캐시를 기본값으로 탑재하고 있다. 일부 최적화된 배포 환경에서는 INT8, 심지어 INT4까지 이미 내려간 사례도 존재한다.

그렇게 놓고 보면, 실전에서 TurboQuant의 실질적 추가 이득은 16비트 → 3비트의 "6배"가 아니라, 8비트 → 3비트, 즉 약 2.7배 압축에 불과할 가능성이 높다. 논문이 주장하는 6배의 절반에도 못 미치는 수치다. 실제로 모든 프론티어 AI 랩은 이 논문이 출판되기 전부터 이미 자체적으로 KV 캐시를 최대한 압축하고 있었다. 시장의 방향성 측면에서 이건 전혀 새로운 것이 아니다

더 강력한 비교 대상도 존재한다. **중국의 DeepSeek가 MLA(Multi-head Latent Attention)라는 혁신적인 구조로 KV 캐시를 무려 28배 압축한 것은 이미 작년 초 DeepSeek R1이 나왔을 때 널리 알려진 사실이다. MLA는 아키텍처 수준의 변경이고 TurboQuant는 기존 모델에 사후적으로 적용 가능하다는 차이가 있으나, 순수한 압축 비율만 놓고 보면 TurboQuant의 "6배"(실질 2.7배)는 MLA의 28배 앞에서 초라하다**.

게다가, 메모리 절감이 발생할 때마다, 우리는 더 큰 컨텍스트 윈도우를 서빙할 수 있는 더 나은 모델을 얻게 된다는 것을 유념해야 한다. 100만 토큰 컨텍스트 윈도우를 가진 모델과 더 짧은 컨텍스트 모델 사이의 품질 차이는 엄청나다. 이것은 사용해본 경험자들은 똑같이 느낄 것이다. 이번에도 다를 바 없다. 사용량이 더 폭발적으로 성장하고, HBM 수요는 추론 쪽에서조차 계속 증가할 것이다.

따라서, TurboQuant의 숫자들이 곧바로 "메모리 슈퍼사이클이 끝난다"는 식의 결론으로 이어지는 것은 전혀 다른 문제다. 따라서 TurboQuant의 진짜 의미는 "메모리 전체를 6배 절감한다"가 아니라, "추론에서 가장 비싼 병목 하나를 추가로 2~3배 더 효율화할 가능성을 열었다"에 가깝다. 이 차이를 이해하지 못하면, 기술을 너무 과소평가하거나 반대로 너무 과대해석하게 된다.

(4) 그래도 의미 있는 지점: "생각보다 빠른 배포가 가능"

그리고 TurboQuant에서 공개한 벤치마크는 상대적으로 아직 작은 모델 중심이고, 70B+ 대형 모델이나 MoE 구조, 초장문 문맥에서 동일한 무손실 압축이 재현되는지는 더 지켜봐야 한다. 또한 "8배 가속"이라는 표현도 특정 어텐션 연산 구간에 한정된 결과이지, end-to-end 전체 추론 처리량 8배를 뜻하는 것은 아니다.

이렇듯 검증 범위가 제한적이긴 해도, 경제적 함의가 있다. 이미 AI 커뮤니티에 있는 개발자들은 어느 정도의 구현을 통해 이 기술의 가능성을 실제로 보여주고 있다.

한 개발자는, 애플의 MLX 프레임워크를 통해 Qwen3.5-35B-A3B 모델에 TurboQuant를 직접 구현해 본 결과, '8.5K, 32.7K, 64.2K 컨텍스트 길이' 모두에서 정확도 손실 없이 캐시 용량을 최대 4.9배까지 줄여냈다고 증언했다.

여기서 추가적인 파인튜닝이나 재학습이 필요 없다는 점이 포인트다. 이는 배포의 장벽이 생각보다 낮다는 사실은, 이 기술이 연구실 안에만 머물 확률보다 실제 추론 스택으로 침투할 확률이 높다는 뜻이다.

더 극적인 실전 구현 사례도 등장했다. 한 개발자가 TurboQuant를 vLLM(오픈소스 서빙 프레임워크)에 직접 구현한 결과, USB 충전기 크기의 소형 기기에서 NVIDIA GB10 위에 4,083,072개의 KV 캐시 토큰을 올리는 데 성공했다. 400만 토큰이면 책 수십 권 분량의 컨텍스트를 주머니 크기의 디바이스에서 유지할 수 있다는 뜻이다.

여기서 흥미로운 통찰은, 이러한 압축 기술이 데이터센터 내부의 TCO 절감에서 끝나지 않을 수 있다는 점이다. 다시 말해 **3비트 수준의 KV 캐시 압축은 모바일 장문맥 추론과 온디바이스 AI의 현실화를 앞당길 수 있다. 이는 AI가 단지 더 싸게 돌아가는 데 그치지 않고, 더 많은 디바이스에 깔리며 배치 가능한 표면적 자체가 넓어진다는 뜻이다**.

그 확산은 다시 더 많은 생성, 더 많은 동기화, 더 많은 저장 수요를 부르며 전체 메모리 스택의 장기 TAM을 키우는 방향으로 이어진다. 결국, 제본스의 역설이다.

3. 치명적 오해의 정정

(1) 공짜 점심은 없다: 메모리를 아끼는 대신 연산을 더 요구한다

TurboQuant의 무작위 회전과 극좌표 변환인 PolarQuant는 소프트웨어적으로 매우 아름다운 수학으로 평가받지만, 이를 물리 세계인 반도체 칩 입장에서는 이야기가 다르다. 이 압축을 풀고 계산하는 과정이 까다로운 '수학 세금(Math Tax)'으로 다가오기 때문이다.

3비트 혹은 4비트로 압축된 데이터를 실제 연산에 쓰려면, 결국 그 데이터를 읽고, 해석하고, 일부는 재구성한 뒤 연산에 투입해야 한다. 이 과정은 공짜가 아니다. 즉, TurboQuant는 단순히 용량을 줄이는 기술이 아니라, 메모리 트래픽 일부를 연산 비용으로 치환하는 기술이다.

특히 범용 GPU 입장에서는 더욱 힘든 일이다. 엔비디아의 GPU는 수천~수만 개의 만능 작업자(CUDA 코어)를 꽉꽉 채워 넣은 것이니, 이 범용 작업자들이 TurboQuant로 압축된 데이터를 처리하려면 '읽기-계산하기-쓰기(Read-Math-Write)'라는 사이클을 거쳐야 한다. 처음 보는 생소한 3비트 데이터를 16비트로 부풀린 다음, "무작위 회전과 극좌표 해독"을 수행하고, 이후에 다시 쓰는 과정이 1초에 수백만 번 반복돼야 한다. 이는 트랜지스터들의 스위칭을 폭증시키고, 결국 국지적 발열인 Hot spot들이 생기는 약점이 된다.

(3) 팩트: TurboQuant는 엔비디아에서도 작동한다

그렇다고 해서 이것을 **구글 TPU의 승리, 엔비디아 GPU의 패배로 읽지는 말아야 한다. TurboQuant는 구글 TPU만 가진 전용 기술이 아니다. 하드웨어 비종속적(Hardware-agnostic)이다**. 애초에 논문의 핵심 벤치마크 자체가 엔비디아 H100 GPU에서 수행되었으며, 4비트 TurboQuant가 H100에서 32비트 대비 8배의 속도 향상을 달성한 것이 실험 결과의 주요 헤드라인이기 때문이다.

TurboQuant가 엔비디아 H100에서도 8배 속도 향상을 달성했다는 사실은, 엔비디아 칩들도 이 세금을 내면서도 충분한 이익을 얻고 있다는 뜻이다. 이러한 사실은 두 가지 차원에서 중요하다.

첫째, On-chip SRAM 계층에서의 이득이 있다. KV 캐시가 3~4비트로 줄어들면 HBM 용량만 절약되는 것에 그치지 않는다. 칩 내부의 L1/L2 SRAM 캐시에 KV 데이터를 훨씬 더 많이 올릴 수 있게 된다. 엔비디아가 Groq LPU에 이 아이디어를 어떻게 활용할지 벌써 상상이 든다. 즉, TurboQuant의 이득은 "용량 절약"만이 아니라 메모리 계층 구조(Memory hierarchy) 전체에서의 성능 향상이라는 점이 중요하다. 결국 TurboQuant이 HBM을 덜 쓰게 만들지 않을 것이라는 게 포인트다.

둘째는 생태계 확장성이다. TurboQuant가 엔비디아 생태계에서도 동일하게 작동한다는 사실은, vLLM·SGLang 같은 오픈소스 서빙 프레임워크에 통합되면 Azure, AWS, 오라클 클라우드 등 모든 AI 서비스 환경에 즉시 배포될 수 있음을 의미한다. 따라서 구글의 TPU v7이 TurboQuant를 '더 효율적으로' 소화하는 것은 맞지만, 엔비디아 GPU가 이 기술 앞에서 '무력화된다'는 해석은 과장일 수 있다.

(3) 구글 TPU v7 (Ironwood)와 실리콘의 역설

물론, 구글이 Ironwood TPU를 설계할 때 TurboQuant을 감안해서 이러한 수학 세금을 줄이는 방향으로 칩의 구조를 짠 것은 사실로 보인다. TPU는 시스톨릭 어레이(Systolic Array) 구조이기 때문에 구조적으로 GPU와 달리 압축 해제와 곱셈 연산이 빠르다.

그리고 가장 재밌는 점은, 구글이 칩을 바라보는 방식이 우리 상식과는 다르다는 점이다. 기업이나 개인이 엔비디아 GPU를 사는 경우, 보통은 한 번 들여온 장비를 5~7년 정도 안정적으로 오래 쓰는 것이 중요하다. 장애가 나면 직접 감당해야 하고, 교체 비용도 크기 때문이다. 그래서 장비를 어느 정도 여유 있게 운용하고, 지나치게 높은 활용률로 몰아붙이지 않는 경우가 많다.

하지만 구글 데이터센터 내부에서 쓰이는 Ironwood 같은 TPU는 성격이 다르다. 구글에게 중요한 것은 개별 칩 하나의 긴 수명이 아니라, 데이터센터 전체가 일정 시간 동안 얼마나 많은 추론과 연산을 처리하느냐다. 다시 말해, 구글은 칩을 "오래 아껴 쓰는 자산"이라기보다, 전체 시스템 안에서 최대 처리량을 뽑아내는 소모품에 가깝게 본다. 이 차이는 운영 방식에서 그대로 드러난다.

구글은 데이터센터 전체를 직접 통제하기 때문에, 칩 하나가 과열되거나 수명이 짧아져도 시스템 전체가 멈추지 않게 설계한다. 즉, 칩 하나를 최대한 오래 쓰는 것이 아니라, 필요하면 더 강하게 혹사시키고 더 빨리 교체하더라도 전체 시스템 생산성을 높이는 편이 낫다고 보는 것이다. TurboQuant와의 관계도 바로 여기서 연결된다.

TurboQuant는 겉으로 보면 메모리를 절약하는 기술처럼 보이지만, 실은 더 정확히 말해 메모리 병목 일부를 계산 부담으로 옮기는 기술이다. KV 캐시를 더 작게 저장할 수 있게 해주지만, 그 대신 압축된 데이터를 읽고, 해석하고, 일부는 복원해서 다시 계산에 투입해야 한다. 즉, 메모리를 덜 쓰는 대신 칩이 더 바빠진다. 그런데 구글에게는 이게 반드시 불리한 것이 아니다. **왜냐하면 지금 AI 인프라의 핵심 병목은 "계산 유닛이 너무 한가해서 남아도는 것"이 아니라, 오히려 메모리 때문에 충분히 돌리지 못하는 것에 가깝기 때문이다. TurboQuant는 이 병목을 낮춰서, 원래라면 메모리 제약 때문에 놀고 있던 연산 자원을 더 오래, 더 빽빽하게 쓰게 만들어준다.**

결국 구글이 칩을 더 공격적으로 운용할 유인이 커지는 이유도 여기에 있다. TurboQuant 같은 기술이 들어오면, 같은 HBM 용량 안에서 더 많은 에이전트 호출을 처리할 수 있다. 그러면 시스템 전체 처리량은 올라가고, 칩의 활용률도 더 높아진다는 점이 포인트다. 물론 그만큼 발열과 전력 부담, 마모 압력도 커질 수 있지만, 이러한 추가 부담은 그리 문제가 되지 않는 셈이다. 그 결과 칩 활용률 상승과 교체 주기 단축, 나아가 반도체 수요의 구조적 확대로 이어질 수 있다. 이것이 바로 이번 사태에서 말하는 '실리콘의 역설'이다.

4. 제본스의 역설 2.0

(1) 틱톡을 낳은 것도 제본스 덕분(?)

제본스의 역설은 19세기에 윌리엄 제본스라는 경제학자가 석탄 효율이 좋아진 증기기관을 보고 깨달은 법칙이다. "자원을 사용하는 효율이 좋아져서 비용이 싸지면, 사람들은 그 자원을 덜 쓰는 게 아니라 오히려 미친 듯이 더 많이 쓴다."

실제 역사적 사례도 숱하게 많다. 비디오 압축 코덱인 “H.265”가 도입되어 10GB짜리 원본 영상이 40MB(260배 압축)로 줄어들었을 때, 인류가 데이터센터 스토리지를 덜 샀는가? 정반대다. 유튜브 4K 영상이 쏟아지고, 틱톡이 탄생하고, 인스타그램 릴스와 유튜브 숏츠가 세상을 뒤덮었다. 결국 인류는 과거보다 수천 배 더 많은 데이터센터/스토리지를 건설했다.

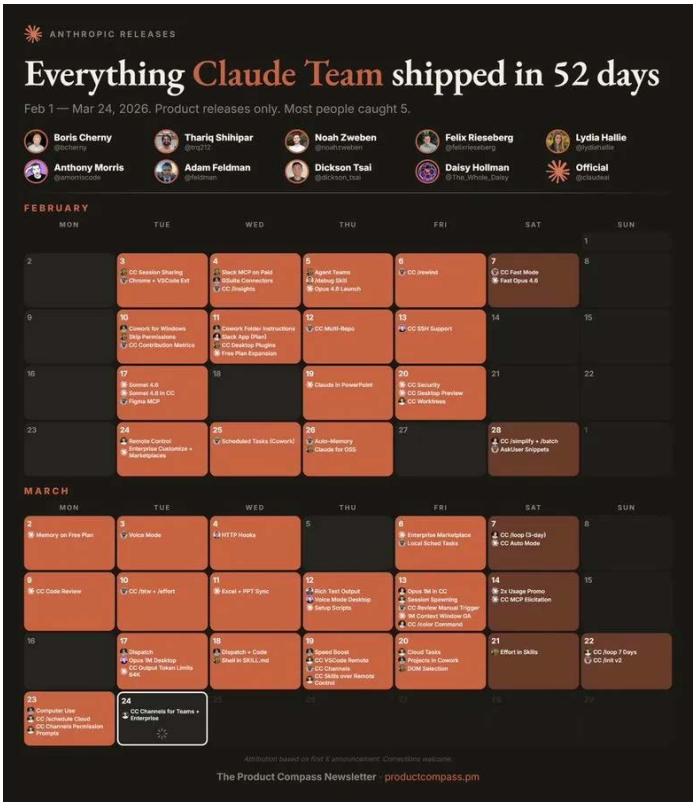
(2) Sora의 죽음의 의미: 2026년 3월, 컴퓨팅은 여전히 부족하다

비디오 코덱의 역사까지 볼 필요도 없다. 제본스 역설의 가장 생생한 증거가 2026년 3월, 바로 이 순간 벌어지고 있기 때문이다. 현재, AI 인프라 시장에서 동시다발적으로 일어나고 있는 세 가지 사건을 직시해야 한다.

첫째, 앤트로픽(Anthropic)의 폭발적 성장이 GPU 공급 한계를 드러내고 있다.

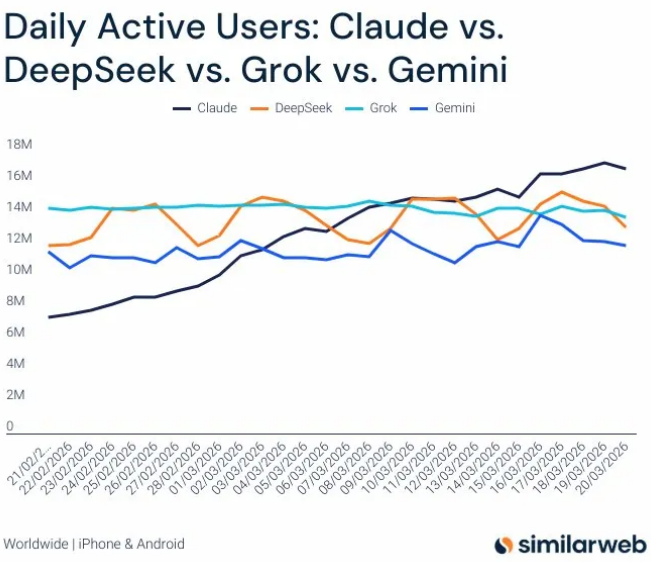
클로드(Claude)는 주간 단위로 30% 성장하고 있다. DAU(일일 활성 사용자)는 2026년 2월 약 600만 명에서 3월 약 1,600만 명으로 불과 한 달 만에 거의 3배 폭증했다. 52일 동안 쏟아낸 제품 업데이트의 양은 타의 추종을 불허한다: Claude Code, Cowork, Computer Use, Voice Mode, MCP 연동, 1M 컨텍스트 윈도우 GA(일반출시)까지 있다.

그림 4. Anthropic이 단 52일 동안 엄청난 속도로 쏟아낸 제품들
실제 사용자 수요를 전제로 빠르게 확장… 추론 인프라는 절감의 대상이 아니라, 제품 확장의 연료



자료: The Product Compass Newsletter, 미래에셋증권 리서치센터

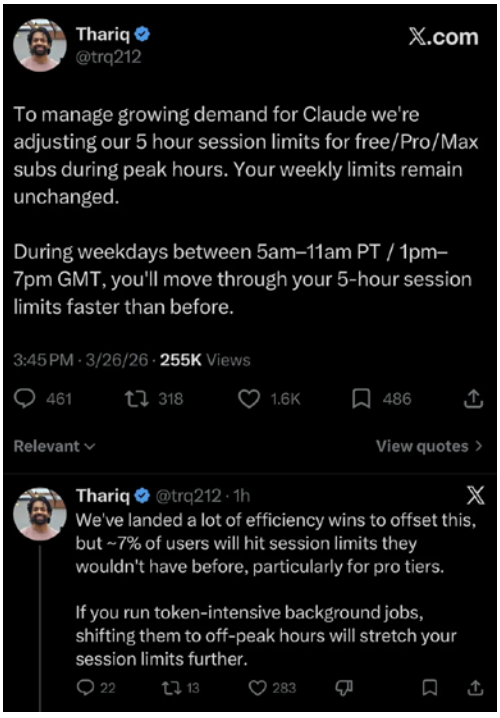
그림 5. Claude vs DeepSeek vs Grok vs Gemini DAU 차트
경쟁 서비스 대비 Claude가 최근 구간에서 가장 강한 상승세



자료: similarweb, 미래에셋증권 리서치센터

그런데 이 성공의 이면에서 벌어지고 있는 일이 결정적이다. 클로드 코드 담당 엔지니어 Thariq가 직접 공개한 증언은, "늘어나는 수요를 관리하기 위해 Free/Pro/Max 구독의 5시간 세션 리밋을 피크 시간대에 조정합니다. 효율화를 많이 달성했지만, 약 7%의 사용자가 이전에는 부딪히지 않았을 세션 리밋에 도달하게 됩니다"라는 내용이었다. AI 추론 인프라가 수요를 따라가지 못하고 있다는 최전선의 최신의 비명이다.

그림 6. Anthropic이 직접 수요가 너무 빠르게 늘고 있어 제한을 조정한다고 말하는 직원의 트윗
"효율성 개선을 많이 해놨지만, 그래도 일부 사용자는 이전보다 더 빨리 한도에 도달할 것"이라 언급



자료: X(@Thariq), 미래에셋증권 리서치센터

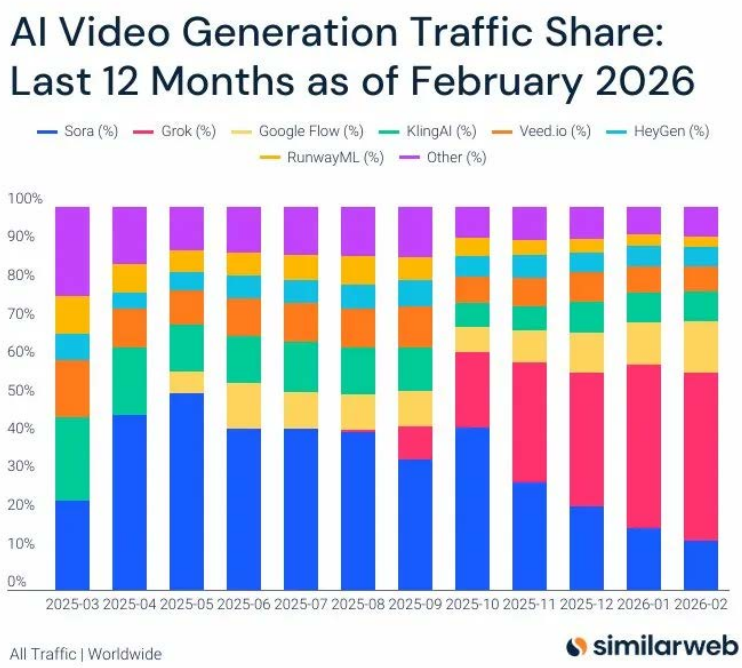
둘째, OpenAI가 Sora를 포기하면서까지 GPU를 재배치했다.

OpenAI는 Sora 비디오 생성에 하루 1,000만~1,500만 달러, 즉, 연간 약 55억 달러의 GPU 비용을 태우고 있었다고 알려졌다. 그리고 며칠 전, Sora에 쓰이던 이 거대한 GPU 풀을 이제 ChatGPT와 새로운 코딩 에이전트 Codex, 그리고 차세대 모델인 "Spud(코드명)"의 사전학습으로 전면 재배치하기 위해 Sora 앱과 API를 공식 종료했다.

이 결정의 무게를 이해해야 한다. 우선 OpenAI의 Sora가 떠난 자리에 xAI의 Grok Imagine이 빠르게 왕좌를 차지하고 있다. SimilarWeb 데이터에 따르면, Sora 트래픽이 급감하는 동안 Grok의 점유율이 수직 상승한 것을 알 수 있다.

그런데 문제는 AI 비디오 생성 시장의 TAM(Total Addressable Market)은 생각보다도 더 거대할 수 있다는 점이다. 대중교통을 타본 사람들은 알겠지만 정말로 많은 사람들이 이 도파민의 원천으로 숏폼 동영상을 아래위로 열심히 스크롤 할 수 있기 때문이다. 샘 알트만이 이러한 거대한 시장을 가만히 내주고 볼 사람이 아니다. 그럼에도 불구하고 Sora를 죽인 것은, 컴퓨팅 자원이 그만큼 절대적으로 부족하기 때문이다. 비디오 생성보다 코딩 에이전트와 차세대 프론티어 모델에 컴퓨팅을 쏟는 것이 생존에 더 급하다는 판단일 것이다.

그림 7. 지난 12개월 동안 AI 비디오 트래픽 주도권이 Sora에서 Grok Imagine 등으로 빠르게 이동



자료: similarweb, 미래에셋증권 리서치센터

한편, Sora 출시 후 103일 만에 종료한 것을 포함해 OpenAI의 피벗 행렬은 이번이 처음이 아니다. 샘 알트만이 "최후의 수단"이라 했던 광고를 16개월 만에 도입한 것, 인앱 쇼핑 기능을 출시하기로 한 것을 취소한 것, 첫 하드웨어 기기 출시를 2027년으로 지연한 것 등이 있다. 이 모든 피벗의 공통 분모는 하나다. 컴퓨팅 자원과 이를 위한 자금이 절대적으로 부족하기 때문에 우선순위를 끊임없이 재조정해야 하는 현실이다.

셋째, TurboQuant가 이 방정식에 던지는 질문.

만약 TurboQuant 같은 효율화 기술이 정말로 추론 비용을 유의미하게 낮춘다면, 그 결과는 OpenAI가 메모리를 덜 사고 돈을 아끼게 될까? TurboQuant 출시로 숨통이 많이 트였다면 곧바로 샘 알트만은 Sora를 유지했을 것이다. 막대한 시장 기회를 인지하면서도, 애초에 비용 때문에 우선순위에 밀려 포기한 것이기 때문이다.

그런데 현실은 어떤가? 사실 TurboQuant 논문이 공개된 지 11개월이 지났는데도 OpenAI는 Sora를 포기할 수밖에 없었다. 즉, 효율화 기술이 이미 존재하고 있는데도 컴퓨팅 수요를 공급이 따라가기 벅차다는 말이다. 이것이 현재 시점의 가장 직접적인 증거다.

따라서, 앞으로도 더 많은 역설을 이끌어낼 기술이 나와야 한다. 효율화는 과거에 비용 제약 때문에 포기했던 거대한 워크로드를 다시 실행 가능하게 만든다. 비디오 생성, 네이티브 멀티모달 에이전트 같은 것들이 앞으로 출현해야 한다는 말이다. 과거에는 GPU가 부족해서 우선순위에서 밀려났던 이 모든 것들이 효율화의 열매를 따는 순간 동시에 부활하여, 총 컴퓨팅 수요를 효율화 이전보다 더 폭발적으로 끌어올린다. 제본스의 이론으로 꽃 피울 수 있는 AI 대중화의 길은 아직도 갈 길이 멀다는 것을 유념하자.

(3) 베테랑의 증언: "모든 양자화를 다 써봤지만, 메모리는 여전히 병목이다"

TurboQuant의 등장은 메모리 수퍼사이클의 종말이 아니라, AI 시장의 진정한 대중화를 가속하는 기폭제다. 추론이 싸지면 에이전트가 폭증하고, 에이전트가 폭증하면 생성 데이터가 폭증하고, 생성 데이터가 폭증하면 스토리지 수요가 폭증한다. 이 연쇄 반응의 어느 단계에서도 메모리 수요가 '줄어드는' 지점을 찾아보기는 힘들다.

일전에 설명했듯 DeepSeek MLA가 28배를 압축했어도 "압축이 메모리 수요를 줄였다"는 결과를 본 적은 없다. 이러한 역사적인 경험만에 더해, 더욱 결정적인 증거는 역시 현장에서 찾을 수 있었다. 10년 가까이 AI 추론 인프라를 직접 구축해 온 한 엔지니어의 증언은 다음과 같다.

"저는 거의 10년 전부터 AI 추론 인프라를 구축해 왔고, 메모리가 용량과 대역폭 양 측면에서 추론 시 가장 큰 병목이라는 인식이 있었습니다. 저는 과거에 모델의 dense 아키텍처와 임베딩 테이블의 메모리 풋프린트를 줄이기 위해 온갖 양자화 접근법을 전부 다 써봤습니다. 그런데도 여전히, 이 오랜 시간이 지난 지금도, 메모리가 여전히 병목입니다."

이 증언의 파괴력은 학술 논문 열 편보다 강하다. 10년간 실제 프로덕션 환경에서 INT8, INT4, 가지치기(Pruning), 임베딩 테이블 압축, 희소 행렬 최적화 등 양자화의 모든 무기를 전부 동원해 본 사람이 내린 결론이다. 매번 양자화로 메모리를 줄일 때마다, 그 여유 공간은 더 큰 모델, 더 긴 컨텍스트, 더 많은 동시 사용자로 즉시 채워졌다는 말이다. 이는 제본스의 역설이 학술 이론도, 투자자들의 희망회로가 아니고, 산업의 당사자가 인정한 구조적 현실이라는 점이 중요하다.

(4) 치명적인 오해: '추론'과 '학습'의 차이

이번에 다시 모습을 드러낸 많은 AI 회의론자들이, 인지하지 못한 결정적인 팩트가 있다. TurboQuant가 6배 압축하는 'KV 캐시'는 철저하게 AI가 답변을 만들어내는 '추론' 단계에서만 쓰이는 메모리라는 점이다. 하지만 메모리 반도체(HBM)의 진짜 거대한 수요는 AI 모델을 가르치는 '학습(Training)' 단계에서 발생한다.

추론(Inference)은 이미 똑똑해진 AI에게 질문을 던지고 답을 받아내는 과정이고, 순전파(Forward Pass)만 발생한다. 그리고 "지금까지 무슨 대화를 나눴지?"를 기억하기 위한 임시 메모장이 KV 캐시다. TurboQuant는 이 임시 메모장을 3~4비트로 구겨 넣는 기술이다.

반면, 학습(Training)은 백지상태의 AI에게 전 세계의 인터넷 데이터를 집어넣고 지능을 만들어내는 과정이다. 순전파뿐만 아니라 역전파(Backward Pass)라는 수학적 뇌수술이 동시에 일어난다.

문제는, 100B 크기의 모델 하나를 훈련시키기 위해 HBM 안에는 4가지의 거대한 데이터 덩어리가 상주해야 한다는 점이다. KV 캐시 정도가 아니다.

① **가중치(Weights): 200GB**. 1,000억 개 × 16비트(2바이트). 3비트로 압축하면 AI는 복잡한 개념을 구분하지 못하고 "백치"가 된다.

② **기울기(Gradients): 200GB**. 가중치 1개당 기울기 1개가 1:1로 매칭. 기울기 값은 0.00000012처럼 극한의 소수점이다. 3비트로 압축하면 이 값은 그냥 '0'으로 반올림되기 때문에, 학습이 영원히 정지하는 문제가 생긴다. 이것이 기울기 소실이다.

③ **옵티마이저 상태(Optimizer States): 1,200GB(1.2TB).** 대중이 거의 알지 못하는 영역이다. 가장 유명한 AdamW라는 옵티마이저는 가중치 1개당 “32비트의 마스터 가중치(4바이트) + 32비트 모멘텀(4바이트) + 32비트 분산(4바이트) = 12바이트”를 요구한다. 이렇게 되면 1,000억 × 12바이트 = 1.2TB. 파라미터 크기의 무려 6배에 달하는 메모리가 오직 '옵티마이저' 하나 때문에 증발한다.

** 모멘텀은 과거 기울기들의 '평균적인 방향과 속도'를 32비트로 저장. AI가 이상한 데이터를 먹고 엉뚱한 방향으로 튀려고 할 때, 이 관성이 "원래 가던 길로 가라"면서 올바른 정답(Global Minima)을 향해 밀어붙임.*
** 분산은 각 가중치의 변동성을 32비트로 측정하여, 널뛰는 가중치에게는 훈련의 효과의 보폭을 줄이고, 움직임이 굼뜬 가중치에게는 보폭을 키워주는 역할.*

④ **활성화(Activations): 수 TB에 달함**. 역전파를 하려면 순전파 때 모든 층에서 발생한 중간 계산 결과물을 전부 HBM에 들고 있어야 한다.

이 모든 것을 합산하면, “가중치(200GB) + 기울기(200GB) + 옵티마이저(1,200GB) = 최소 1.6TB + 활성화 수 TB”가 된다. 1조 개의 파라미터를 가진 LLM이라면 이 필요 메모리 규모는 당연하게도 수백 테라바이트로 폭증한다.

TurboQuant의 압축 방식이 여기에 개입할 수 없는 구조적 이유는 명확하다. 인공지능의 신경망은 수백 개의 층과 수천억 개의 노드가 거미줄처럼 얽혀 끝없이 곱셈을 반복하는 비선형적인 구조다.

입력층의 초기 가중치에서 0.000001이라는 미세한 오차가 발생하면, 다음 층으로 넘어갈 때마다 기하급수적으로 증폭된다. 로켓 발사 각도가 0.001도만 틀어져도 수억 킬로미터 후에는 목표 행성을 완전히 빗나가는 나비효과와 같다. 이건 추론처럼 어느 정도 손실 압축을 허용할 수 있는 영역이 아니다. 즉, TurboQuant는 추론 메모리 최적화에는 위력적일 수 있어도, 훈련용 HBM 수요의 구조에는 영향을 끼칠 수 없다.

시장을 놀래킨 "6배 압축의 마법"은 완성된 AI를 배포할 때의 배달통 크기를 줄여준 것일 뿐이다. AI의 지능을 주조하는 훈련의 가마솥(HBM) 크기는 1g도 줄어들지 않는다.

5. 구글의 엔드게임: 수조 개의 토큰과 시맨틱 검색

(1) "인터넷 전체를 담아야 한다"

2026년 2월 12일, 구글 AI의 수장 제프 딘(Jeff Dean)은 팟캐스트에서 현재 AI의 한계를 명확히 지적한 바 있다. 현재 업계에서는, 100만~200만개의 토큰 컨텍스트를 밀어붙이고 있지만, 그의 기준에선 "턱없이 짧다"는 것이다. 그가 진짜 원하는 마법은 "수조(Trillions) 개의 토큰"을 AI의 머릿속에 한 번에 집어넣는 것이라고 밝혔다.

이 수조개의 토큰이라는 규모는, AI가 대답할 때 인터넷 전체의 데이터를 실시간으로 전부 뒤져보는 수준이다. 그리고 유튜브의 모든 비디오 픽셀을 동시에 비교 분석하는 것이다. 또한 개인의 평생 기록(이메일, 사진, 문서, 비행기 티켓)을 프라이버시 권한 하에 단일 AI가 전부 통달하는 것이다. 이 정도 규모의 기억력을 감당해야 한다는 게 지금 구글의 생각이다.

그런데, 현재의 트랜스포머 아키텍처는 이차 복잡도(Quadratic Complexity)다. 컨텍스트가 2배 늘면 계산량은 4배가 된다. 수조 개의 토큰을 다루는 것은 지구상의 모든 반도체를 긁어모아도 불가능하다. 그래서 제프 딘은 "수조 개를 다 처리하는 것과 같은 '환상(Illusion)'을 주어야 한다"고 말한다.

(2) TurboQuant는 메모리 절감 기술이면서 무엇보다 검색 인프라 기술

제프 딘의 “수조 개 토큰의 환상”을 현실로 만드는 핵심 열쇠가 벡터 검색(Vector Search)이다. AI는 세상의 모든 단어, 문장, 이미지의 의미를 고차원 벡터(수백 개의 숫자로 이루어진 방향성 데이터)로 변환해서 이해한다. 이 개념이 중요한 이유는, AI 시대 검색의 핵심은 더 이상 키워드 매칭이 아니기 때문이다. 사용자의 질문을 이해하고, 거대한 의미 공간에서 가장 관련성 높은 조각을 찾아 실시간으로 결합해야 한다. 즉, “시맨틱 벡터 검색”이다.

그런데, 기존의 PQ, RabitQ 같은 양자화 기술은 데이터를 일일이 훑어보고 데이터 특성에 맞춘 무거운 코드북을 만들어야 했다. 인덱싱에 며칠씩 걸리는 끔찍한 병목이 존재했다. 그리고 TurboQuant의 무작위 회전 방식이 여기서 진가를 발휘한다. 벡터들을 아주 적은 비트수로 압축하기 때문에, 인덱싱 시간이 '사실상 제로'가 된다는 것이다.

실제로, TurboQuant가 밝힌 벤치마크 점수를 보면, GloVe 데이터셋(d=200) 기준, TurboQuant는 2비트, 4비트로 극단적으로 압축한 상태에서도 1@k Recall이 가장 빠르게 1.0에 도달했다. 쉽게 말해, 기존의 방식인 PQ보다, 이번에 나온 TurboQuant가 속도는 수천 배 빠르면서도 '가장 정확한 답'을 찾아낸다는 말이다.

종합하자면, TurboQuant를 단순한 KV 캐시 절감 기술로만 보면, 구글이 왜 여기에 깊게 매달리는지 설명이 잘 되지 않는다. 더 큰 그림은 검색 패권을 AI로 다시 더 세게 쥐겠다는 선언인 것이고, 이를 위해 인터넷상의 인류의 모든 지식을 실시간으로 검색하는 시맨틱 검색의 물리적 뼈대를 만들고 있는 것이다. 따라서, TurboQuant는 메모리 절감 논문이면서도, 더 중요한 것은 그것이 동시에 검색 인프라 논문이라는 점이다.

구글의 엔드게임은 바로 이것이다.

6. 결론: 흔들림 없는 구조적 슈퍼사이클

(1) 제본스의 역설은 이번에도 작동한다, “대압축은 대확장의 서막이다”

시장이 늘 헷갈리는 지점이 있다. 효율이 좋아지면 총수요가 줄어들 것이라고 자동으로 가정하는 것이다. 그러나 인프라 산업은 정반대로 움직이는 경우가 더 많다. 비용이 낮아지면 사람들은 덜 쓰는 것이 아니라 더 쓴다. 그리고 AI는 그 경향이 특히 극단적이다.

한 시장 참여자의 증언이 이 사태의 광기를 정확히 포착했다. "인터넷 전체가 TurboQuant가 오늘의 거대한 새 돌파구라는 수백 개의 게시물로 도배되었다. 근데 아무도 이게 2025년 4월에 공개된 거라는 걸 눈치 못 챈 것 같은데? 대체 뭘 놓치고 있는 거지?"

사실 그가 놓친 것은 없다. 시장이 놓치고 있었던 것이다.

흥미로운 점은, 프론티어 AI 연구소(Google DeepMind, OpenAI, Anthropic)들이 공유하는 일종의 침묵의 룰이 존재한다는 점이다. 절대, **자신들의 공개 프로덕션 모델(Gemini, GPT)의 각자의 내부 핵심 경쟁력을 결정짓는 '제로 투 원'의 마법은 논문으로 출판하지 않는다.** 그러니까, 구글이 TurboQuant를 논문으로, 그것도 상세한 증명과 함께 세상에 공개했다는 사실 자체가 지니는 함의는 명확하다.

첫째, **KV 캐시 양자화는 대부분 연구소들이 이미 다루고 있는 '점진적 개선'의 영역으로 전락했다.** 둘째, 구글 딥마인드 내부의 차세대 팀은 이미 3비트 양자화를 넘어, 상태 공간 모델(SSM)의 선형적 특징을 트랜스포머에 융합하거나, PIM 같은 완전히 다른 차원의 물리적 패러다임으로 진입했을 가능성도 있다.

이 맥락을 더욱 견고하게 뒷받침하는 증거 중 하나는, TurboQuant의 핵심 기술인 PolarQuant의 기초를 닦은 원 연구자 자체가 구글을 이미 2년 전에 떠난 점이다. 핵심 인재가 퇴사한 뒤에야 연구가 결실을 맺고, 1년 전에 논문으로 공개되고, 또 1년이 지나서야 시장이 반응했다. 이러한 시차(연구 → 논문 → 시장 반응)야말로, TurboQuant가 프론티어 랩들의 최전선에서는 이미 '소화 완료된 기술'임을 웅변한다.

(2) 메모리 슈퍼사이클: 무엇이 바뀌지 않는가

이번 이슈로 바뀌는 것은 추론 비용 곡선이다. 긴 문맥과 멀티턴 대화, 에이전트 워크플로, 시맨틱 검색의 경제성이 좋아진다. 더 많은 사용량이 가능해지고 더 많은 서비스가 열린다. 이것은 분명한 변화다.

그러나 바뀌지 않는 것도 있다. 학습/훈련용 HBM 수요의 본체는 그대로다. HBM 수요와 KV 캐시 수요는 동일하지 않다. 효율화는 총수요를 줄이지 않는다. 시장이 지금 보고 있는 것은 비용 절감의 표면이다. 그러나 밑바닥에서 진행되는 것은, 더 큰 스케일업(더 큰 모델, 더 많은 데이터, 더 많은 컴퓨팅)을 위한 인프라 배관 공사다. 그리고 그 배관 공사의 끝에서 더 강해지는 것은, 아이러니하게도 바로 메모리다.

TurboQuant가 추론 비용을 낮춘다고 해서, 메모리 산업의 전략적 지위가 흐려지는 것이 아니다. 오히려 반대다. 추론 사용량이 늘어나면 HBM은 더 높은 활용률로 더 혹사되고, 수요는 트리클 다운된다. 즉, DRAM은 더 많은 warm context를 떠받치며, eSSD는 더 많은 유저들의 장기 상태 정보 및 로그 등을 담아야 한다. 특히 과거 '콜드 데이터 창고'로 취급받던 eSSD는 AI 연산의 일부를 떠받치는 Active Storage로 격상될 가능성이 높다.

물론 단기적으로는 시장 심리가 흔들릴 수 있다. 압축이라는 단어 하나만으로도 메모리 업종은 먼저 맞는다. 그러나 그 흔들림을 그대로 펀더멘털 훼손으로 읽는 것은 무리다. TurboQuant의 경제적 의미는 "HBM이 덜 필요해진다"가 아니라, 같은 HBM으로 더 많은 추론을 돌리게 만들고, 그 결과 계층형 메모리 전체 사용량을 다 키우는 방향에 가깝다.

TurboQuant 같은 기술이 정말 메모리 수요를 파괴했다면, TurboQuant 논문 발표 이후 지난 11개월 동안 실제 시장과 산업의 숫자는 왜 정반대로 움직였는가? 샌디스크의 경영진은 어째서 이러한 효율화가 오히려 ROI를 높여 수요를 자극할 수 있다고 말하는가? 왜 업계는 더 많은 GPU와 더 많은 HBM, 더 많은 DRAM, 더 많은 SSD를 동시에 요구하고 있는가?

위 질문들에 정직하게 답하면, 결론은 생각보다 단순해진다. 알고리즘의 우아함이 메모리를 파괴하는 것이 아니라는 점이다. 수학적 알고리즘의 우아함은 오히려 세상을 메모리 중독으로 몰아넣는다. 이번 TurboQuant 쇼크는 DeepSeek의 MLA에 이어 두 번째 증명이다.

당장의 현실도 그렇다. **Claude는 성장 속도 때문에 세션 리밋을 조이고 있고, OpenAI는 GPU를 가장 수익성 높은 방향으로 재배치하고 있다. 만약 효율화 기술이 진짜 총수요를 낮췄다면, 이런 종류의 공급 부족과 우선순위 전쟁은 덜 치열해야 했다. 그런데 현실은 정반대다. 효율화 기술이 이미 존재하는데도 컴퓨팅 수요가 공급을 초과하고 있다**.

'효율화는 수요를 죽이지 않는다. 지금까지 비싸서 못 했던 일을 하게 만든다'를 유념하자.

★ P.S. 1년 전의 데자뷰

1년 1개월 전에 작성했던 "DeepSeek 쇼크" 총정리 보고서(2025.2.10)를 되돌아본다.

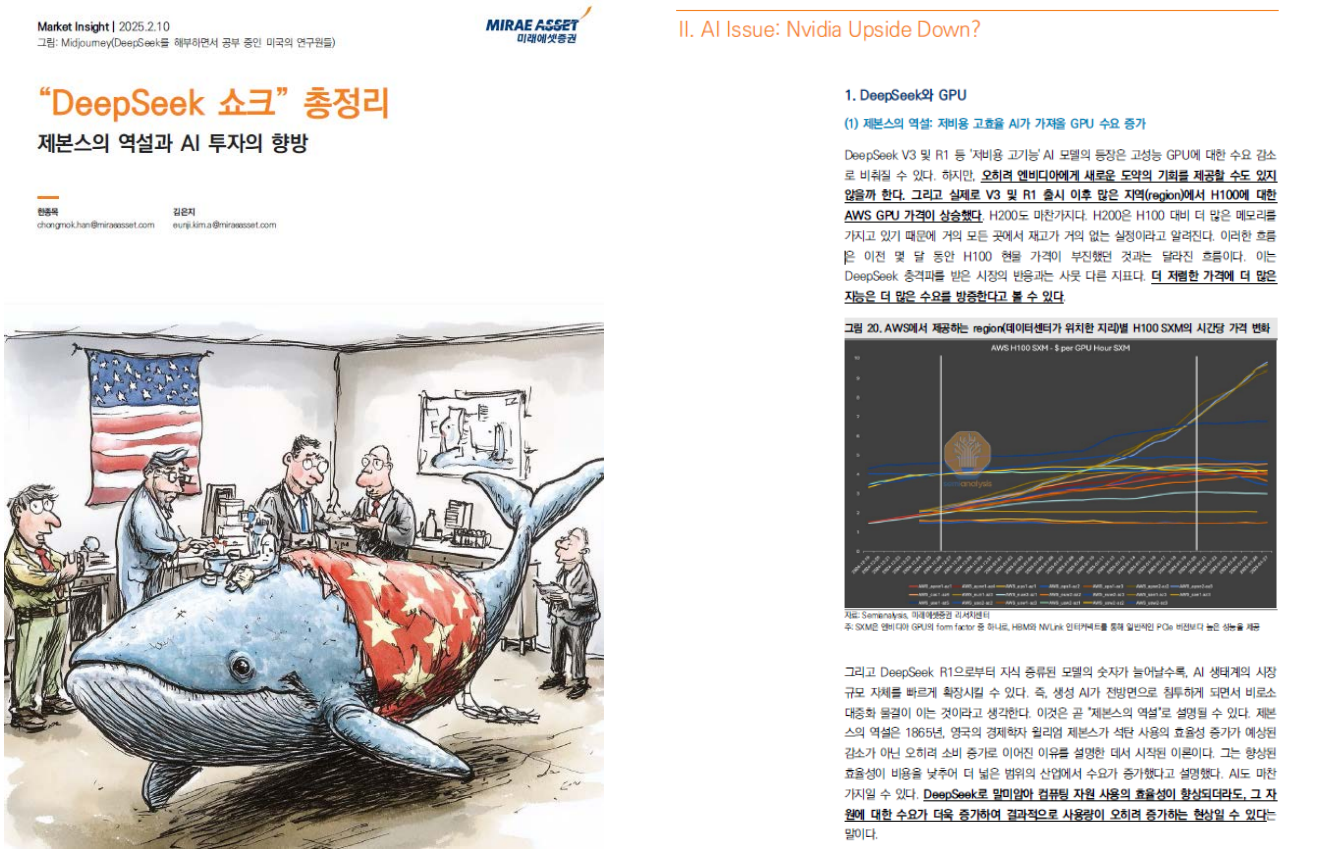
그 때도 동일한 프레임이었다. '저비용 고기능' AI 모델(DeepSeek R1)이 등장하자 시장은 "GPU 수요가 끝났다"며 엔비디아의 시가총액 수백조 원을 날렸다. 우리는 그 날 제본스의 역설을 인용하며 "저비용 고기능 AI가 가져올 GPU 수요 증가"를 주장했고, 사티아 나델라도 같은 날 같은 논리를 트윗했다. "제본스의 역설처럼, AI가 더 효율적으로 바뀜에 따라, AI는 대중화될 것이고 수요는 로켓처럼 올라갈 것."이라는 트윗이었다.

그 후 무슨 일이 벌어졌는가? 하이퍼스케일러들의 CAPEX는 폭증했고, 엔비디아의 주가는 DeepSeek 쇼크 저점에서 회복을 넘어 신고가를 경신했으며, 메모리 반도체 주가는 역사적 랠리를 이어갔다.

지금 TurboQuant 쇼크는 DeepSeek 쇼크의 완벽한 데자뷰다. 기술의 이름만 바뀌었을 뿐, 시장의 반응 패턴, 베어들의 논리 구조, 그리고 그 논리가 부서지는 과정까지 동일하다.

"역사는 운율에 따라 흐른다."

그림 8. DeepSeek 쇼크에 대해서 작년 2월에 작성했던 AI 인덱스 보고서



Market Insight | 2025.2.10

그림: Midjourney(DeepSeek를 해부하면서 공부 중인 미국의 연구원들)

MIRAE ASSET 미래에셋증권

"DeepSeek 쇼크" 총정리

제본스의 역설과 AI 투자의 향방

현종목 chongmok.han@miraeasset.com
김은지 eunji.kim.a@miraeasset.com

II. AI Issue: Nvidia Upside Down?

1. DeepSeek와 GPU

(1) 제본스의 역설: 저비용 고효율 AI가 가져올 GPU 수요 증가

DeepSeek V3 및 R1 등 '저비용 고기능' AI 모델의 등장은 고성능 GPU에 대한 수요 감소로 비춰질 수 있다. 하지만, **오히려 엔비디아에게 새로운 도약의 기회를 제공할 수도 있지 않을까 한다. 그리고 실제로 V3 및 R1 출시 이후 많은 지역(region)에서 H100에 대한 AWS GPU 가격이 상승했다.** H200도 마찬가지다. H200은 H100 대비 더 많은 메모리를 가지고 있기 때문에 거의 모든 곳에서 재고가 거의 없는 실정이라고 알려진다. 이러한 흐름은 이전 몇 달 동안 H100 현물 가격이 부진했던 것과는 달라진 흐름이다. 이는 DeepSeek 충격파를 받은 시장의 반응과는 사뭇 다른 지표다. **더 저렴한 가격에 더 많은 자원을 더 많은 수요를 방증한다고 볼 수 있다.**

그림 20. AWS에서 제공하는 region(데이터센터가 위치한 지리)별 H100 SXM의 시간당 가격 변화

자료: Semianalysis, 미래에셋증권 리서치센터

주: SXM은 엔비디아 GPU의 form factor 중 하나로, HBM와 NVLink 인터커넥트를 통해 일반적인 PCIe 버전보다 높은 성능을 제공

그리고 DeepSeek R1으로부터 지식 증류된 모델의 숫자가 늘어날수록, AI 생태계의 시장 규모 자체를 빠르게 확장시킬 수 있다. 즉, 생성 AI가 전방면으로 침투하게 되면서 비로소 대중화 물결이 이는 것이라고 생각한다. 이것은 곧 "제본스의 역설"로 설명될 수 있다. 제본스의 역설은 1865년, 영국의 경제학자 윌리엄 제본스가 석탄 사용의 효율성 증가가 예상된 감소가 아닌 오히려 소비 증가로 이어진 이유를 설명한 데서 시작된 이론이다. 그는 향상된 효율성이 비용을 낮추어 더 넓은 범위의 산업에서 수요가 증가했다고 설명했다. AI도 마찬가지일 수 있다. **DeepSeek로 말미암아 컴퓨팅 자원 사용의 효율성이 향상되더라도, 그 자원에 대한 수요가 더욱 증가하여 결과적으로 사용량이 오히려 증가하는 현상일 수 있다**는 말이다.

자료: 미래에셋증권 리서치센터

Compliance Notice

- 당사는 자료 작성일 현재 조사분석 대상법인과 관련하여 특별한 이해관계가 없음을 확인합니다.
- 당사는 본 자료를 제3자에게 사전 제공한 사실이 없습니다.
- 본 자료를 작성한 애널리스트는 자료작성일 현재 조사분석 대상법인의 금융투자상품 및 권리를 보유하고 있지 않습니다.
- 본 자료는 외부의 부당한 압력이나 간섭없이 애널리스트의 의견이 정확하게 반영되었음을 확인합니다.

본 조사분석자료는 당사의 리서치센터가 신뢰할 수 있는 자료 및 정보로부터 얻은 것이나, 당사가 그 정확성이나 완전성을 보장할 수 없으므로 투자자 자신의 판단과 책임하에 종목 선택이나 투자시기에 대한 최종 결정을 하시기 바랍니다. 따라서 본 조사분석자료는 어떠한 경우에도 고객의 증권투자 결과에 대한 법적 책임소재의 증빙자료로 사용될 수 없습니다. 본 조사분석자료의 지적재산권은 당사에 있으므로 당사의 허락 없이 무단 복제 및 배포할 수 없습니다.