



2026. 3. 27 (금)

착각의 반복, 시간이 지나면 다시 알게 될 것들

- TurboQuant와 같은 KV캐시 효율화 기술은 계속 연구/진보해왔음(KVTC, NVFP4 KV cache 등). TurboQuant가 이번에 처음 제안된 것도 아님('25.4 처음 제안)
- 효율화 기술이 없으면 안될 정도로 폭발적으로 늘어나는 토큰 소비량이 더 중요
- 매크로 악재가 다른 것도 악재로 보이게 하는 '25년 DeepSeek 시기와 닮은 착각

글로벌 투자전략-선진국

Analyst 황수욱

soowook.hwang@meritz.co.kr

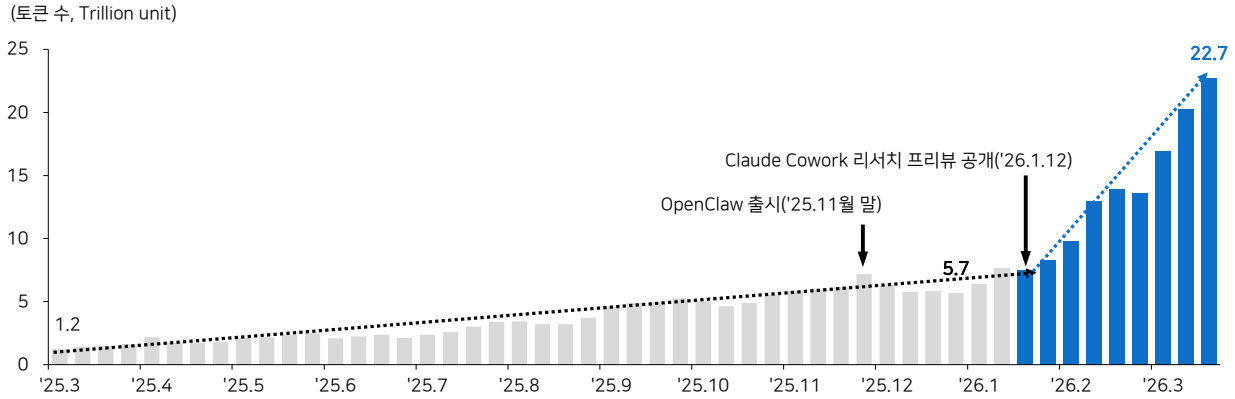
TurboQuant발 주가 반응은 시장의 착각이라는 의견. KV 캐시 효율화 기술은 지속적으로 연구/진보해왔음. 최근 제안되는 관련 기술로는 엔비디아의 NVFP4 KV cache(실행중인 KV cache footprint 축소, KV메모리 50% 절감, TTFT 3배 개선), KVTC(재사용 가능한 KV cache를 압축해 저장, 최대 20배 메모리 수요 절감) 등이 대표적. 구글의 TurboQuant도 메모리 6배 절감 등 구체적인 숫자를 내놓은 점에 시장이 민감하게 반응

지금 시장이 TurboQuant에 주목한 이유는 3월 24일 구글 리서치 블로그에 다시 크게 꺼내들었기 때문. 그런데 이 기술들이 처음 제안된 시점은 최근이 아님. TurboQuant는 2025년 4월 처음 제안되었고, KVTC는 비교적 최신인 2025년 11월, NVFP4 KV cache는 2025년 6월임. 시장 반응이 합리적이었다면, 이런 기술 진보에 대한 주가 반응은 최근이 아니었어야 할 것

합리적이지 않다고 보는 이유는 또다른 이유는 숫자 때문. 설령 어떤 기술이 KV cache 저장용량/처리속도를 6배~20배 개선한다고 해도 AI 에이전트로 연산량(토큰 처리량)이 증가하는 속도가 훨씬 더 빠를 것(그림 1). 최근 확산되는 OpenClaw 등 코드 기반 AI 에이전트는 AI 챗봇 대비 약 만 배 더 많은 토큰 소비(그림 2). KV cache 기술이 AI 하드웨어 수요를 줄이는 게 아니라, 기술이 없으면 안되는 상황으로 인지. 2025년부터 이미 KV cache를 수십 배 줄일 수 있는 기술을 연구하고 있는 엔비디아. 그런데 GTC 2026 기술자 미팅에서 여전히 메모리 기업을 고객사보다 중요하게 여긴다고 한 이유를 생각해볼 필요

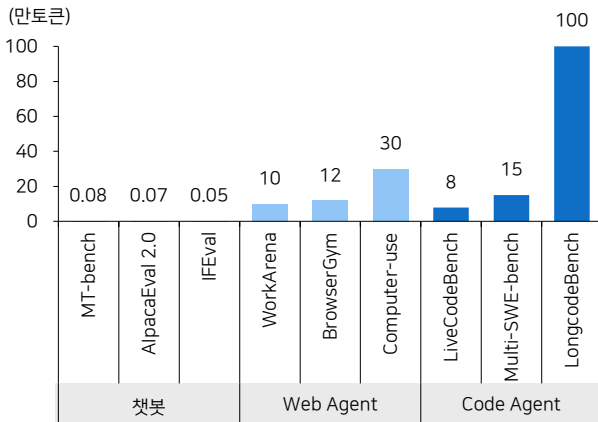
2025년 1~4월과 매우 유사한 상황(그림 4). DeepSeek 기술이 AI 연산량 자체를 줄여 AI 하드웨어 과잉공급이 온다고 우려하던 상황에 무역 분쟁 같은 매크로 변동성이 기존 대비 100배 연산량을 늘리게 한 Reasoning AI의 본질을 보지 못하게 함. 지금도 마찬가지로, 전쟁이 AI 에이전트의 본질을 보지 못하게 하면서 사스포칼립스 우려 등 악재에만 시장이 반응하게 만드는 중. 매크로 변동성이 문제

그림1 OpenRouter 주간 토큰 처리량: AI 에이전트 제품화 이후 폭발적으로 늘어나는 AI 토큰 수요(연산 수요)



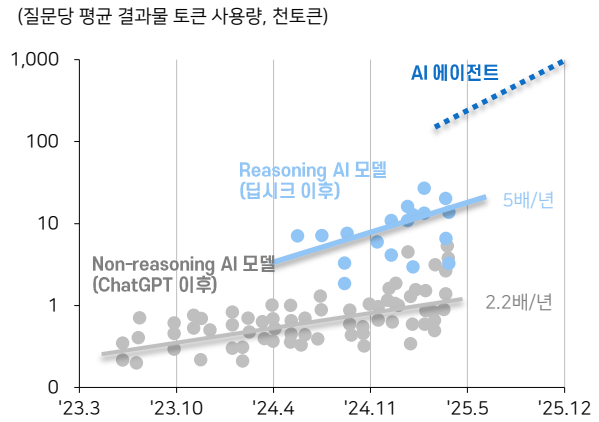
자료: OpenRouter, 메리츠증권 리서치센터

그림2 코드 기반 에이전트는 AI 챗봇 대비 약 만 배 토큰 소비



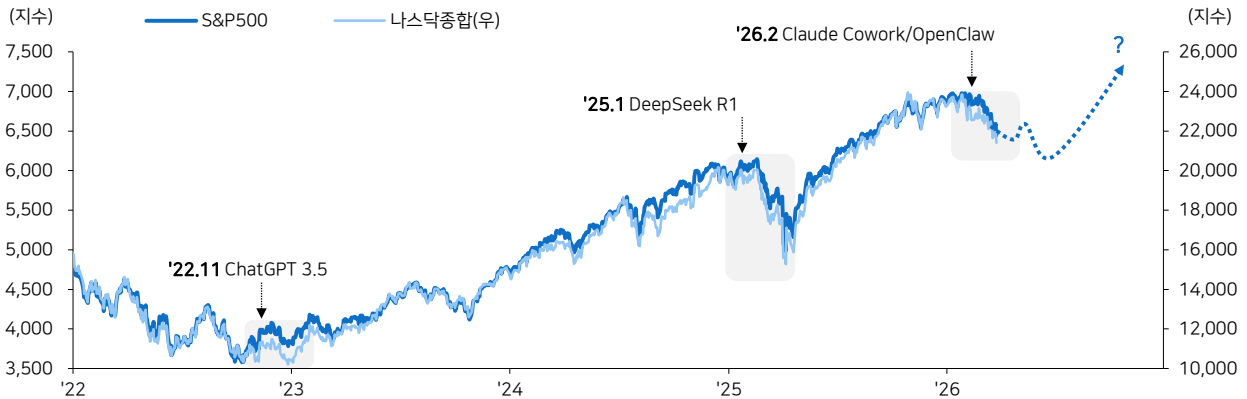
자료: Haoran Wu(2025), 메리츠증권 리서치센터

그림3 OpenClaw는 답시크와 유사한 Regime Shift 요인



자료: Epoch AI, 메리츠증권 리서치센터

그림4 기술 변곡점에 우연하게 다시 찾아온 매크로 변동성: 펀더멘털/기술도 호재는 눈에서 가리고, 악재만 보이게 하는 재료



자료: Bloomberg, 메리츠증권 리서치센터

Compliance Notice

- 본 조사분석자료는 제3자에게 사전 제공된 사실이 없습니다.
- 당사는 자료작성일 현재 본 조사분석자료에 언급된 종목의 지분을 1% 이상 보유하고 있지 않습니다.
- 본 자료를 작성한 애널리스트는 자료작성일 현재 추천 종목과 재산적 이해관계가 없습니다.
- 본 자료에 게재된 내용은 본인의 의견을 정확하게 반영하고 있으며, 외부의 부당한 압력이나 간섭 없이 신의 성실하게 작성되었음을 확인합니다.

본 자료는 투자자들의 투자판단에 참고가 되는 정보제공을 목적으로 배포되는 자료입니다. 본 자료에 수록된 내용은 당사 리서치센터의 추정치로서 오차가 발생할 수 있으며 정확성이나 완벽성은 보장하지 않습니다. 본 자료를 이용하시는 분은 본 자료와 관련한 투자의 최종 결정은 자신의 판단으로 하시기 바랍니다. 따라서 어떠한 경우에도 본 자료는 투자 결과와 관련한 법적 책임소재의 증빙자료로 사용될 수 없습니다. 본 조사분석자료는 당사 고객에 한하여 배포되는 자료로 당사의 허락 없이 복사, 대여, 배포 될 수 없습니다.
