

생성형 AI 모델 이용료 다변화 동향과 시사점

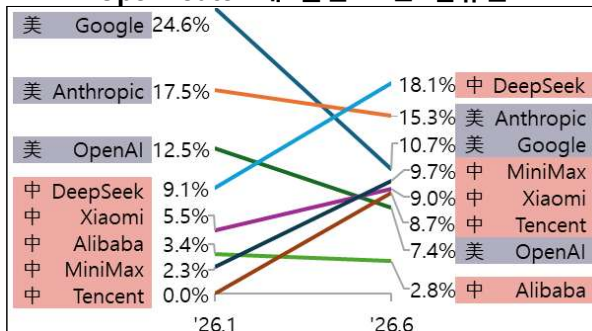
산업기술리서치센터 AI·ICT팀
순수현 (pure.hyun@kdb.co.kr)

- ◆ 기업용 생성형 AI 시장에서 비용 효율성을 중시하는 수요가 확산되면서 저가 모델 비중이 늘고 있으며, AI 모델 개발사는 가격을 세분화한 제품군을 출시하며 대응
- ◆ 가격 경쟁 심화는 개발사의 수익성을 압박할 수 있으나, 동시에 가격 장벽 완화가 시장 성장을 촉진할 수 있어 향후 사용량 확대와 원가 절감이 핵심 과제

□ 기업용 생성형 AI 시장은 비용 효율성 중심의 수요가 확대되는 추세

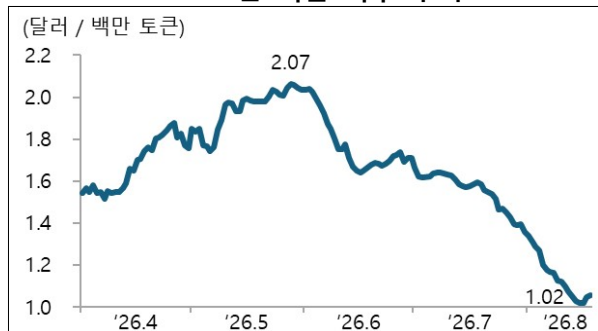
- AI 모델 중개 플랫폼 OpenRouter* 내 월간 토큰** 처리량 순위에 따르면, 초저가 정책의 DeepSeek가 점유율 1위로 올라서는 등 저가 모델 개발사 약진
 - * 기업용 AI 서비스 구축에 활용되는 AI 모델 500여 개를 제공하는 대표적인 중개 플랫폼으로, 천만 명 이상의 고객 보유
 - ** 정보를 AI가 연산할 수 있도록 나눈 데이터 단위
 - 고가·고성능 모델 위주의 미국 기업이 '25.9월 75% 이상 점유했으나 점유율 지속 하락하며 저가 위주의 중국 기업에 추월
 - ※ 美 주요 3사(OpenAI, Anthropic, Google)의 토큰 처리량 점유율: '26.1월 54.6% → '26.6월 33.4%, 中 주요 5사(DeepSeek 등 하단 차트 참고)의 토큰 처리량 점유율: '26.1월 20.3% → '26.6월 48.3%
- 소비자의 토큰당 가중평균 지불 금액을 나타내는 'LLM 토큰 지출 지수'가 최근 하락세를 보이며, 저가 모델 사용 비중이 확대됨을 시사
 - 고가 모델 대비 저가 모델 이용이 빠르게 증가하며 평균 지불액 감소로 지수 하락

OpenRouter 내 월간 토큰 점유율



자료 : OpenRouter 홈페이지 인용

LLM 토큰 지출 지수 추이



자료 : Bloomberg(2026.8), "SDLLMTK"

- AI 도입 기업들은 임직원의 토큰 사용 장려 기조에서 지출 관리 기조로 전환
 - Uber는 직원들의 AI 활용을 장려하고자 토큰 사용량 순위표를 게시하였으나, 4개월 만에 연간 AI 예산을 모두 소진한 뒤 직원 지출 상한제를 도입하였고 안랩, 넷마블, NHN 등 국내 기업들도 지출을 제한하거나 한도 도입을 검토 중

□ 개발사는 점유율 확대를 위해 다양한 가격대의 모델을 출시하며 수요 변화에 대응

- 모델 개발사는 AI 모델의 성능별로 가격을 차등화해 소비자 선택 폭 확대
 - OpenAI, Anthropic은 성능에 따라 고가형·균형형·저가형 모델 출시
 - 한편, Moonshot AI는 전작 대비 3배 이상 이용료를 높인 균형형 모델 'Kimi K3'를 출시하며 저가형 중심의 모델 제품군을 상위 가격대로 확대

주요 모델 개발사의 가격 다변화 예시

(단위 : 달러 / 백만 토큰)

개발사(국가)	모델	지능 지수 ^{주1}	입력 토큰 ^{주2} 비용	출력 토큰 ^{주3} 비용	비고
OpenAI(美)	GPT-5.6 Sol	61	5	30	고가형
	GPT-5.6 Terra	57	2.5 → 2	15 → 12	균형형
	GPT-5.6 Luna	52	1 → 0.2	6 → 1.2	저가형
Anthropic(美)	Claude Fable 5	62	10	50	고가형
	Claude Opus 5	63	5	25	고가형
	Claude Sonnet 5	55	3 → 2	15 → 10	균형형
	Claude Haiku 4.5	30	1	5	저가형
Moonshot AI(中)	Kimi K3	60	3	15	균형형
	Kimi K2.6	45	0.95	4	저가형

주1 : 독립 평가기관 'Artificial Analysis'의 'Intelligence Index' 인용

주2 : 사용자가 모델에 질문을 입력할 때 모델이 질문의 구조와 의도를 파악하는 데 사용되는 토큰

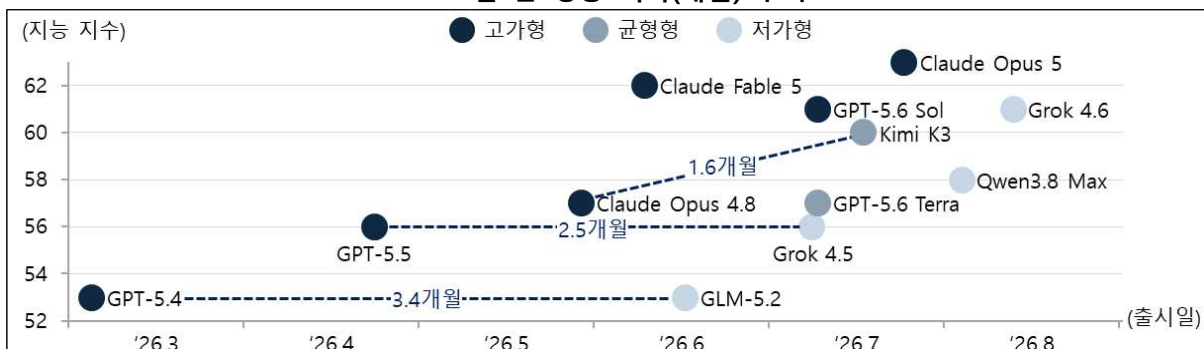
주3 : 모델이 답변을 출력할 때 답변의 내용을 순차적으로 구성하는 데 사용되는 토큰

자료 : Artificial Analysis(2026.8), "Artificial Analysis Intelligence Index v4.1.1" 및 각 사 홈페이지 참고하여 산업은행 재구성

□ 성능이 개선된 저가 모델의 확산으로 모델 개발사 간 가격 경쟁 심화

- 중저가 신모델이 이전 세대 고가 모델 성능 수준에 도달하는 기간이 단축되면서 소비자의 대안 선택 폭은 확대되는 반면, 모델 개발사의 수익성은 악화
 - Anthropic의 'Claude Opus 4.8' 출시 후 단 1.6개월 만에 Moonshot AI가 'Claude Opus 4.8' 대비 성능은 향상되고 이용료는 60% 수준인 'Kimi K3'를 출시
 - OpenAI는 '26.7월 신모델 3종을 동시에 출시하였으며, 출시 3주 만에 균형형 모델 'GPT-5.6 Terra'와 저가형 모델 'GPT-5.6 Luna'의 가격 인하

AI 모델 간 성능 격차(개월) 추이

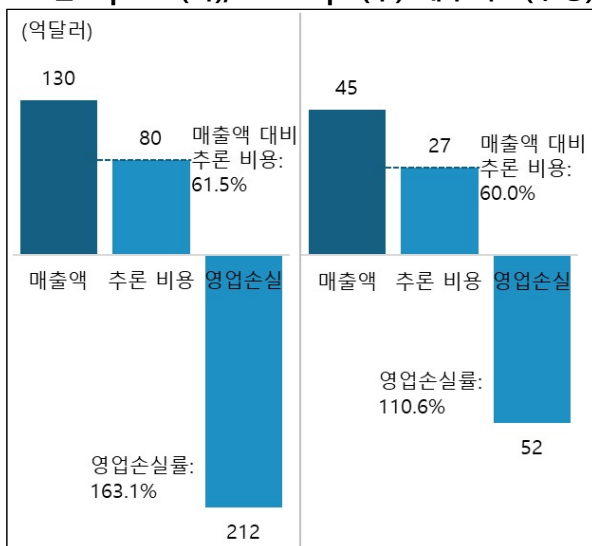


자료 : Artificial Analysis(2026.8), "Artificial Analysis Intelligence Index v4.1.1", 산업은행 재구성

□ 모델 개발사는 가격 경쟁력을 바탕으로 고객 기반 및 토큰 사용량을 확대함과 동시에 수익성 확보를 위한 토큰 원가 절감을 병행할 필요

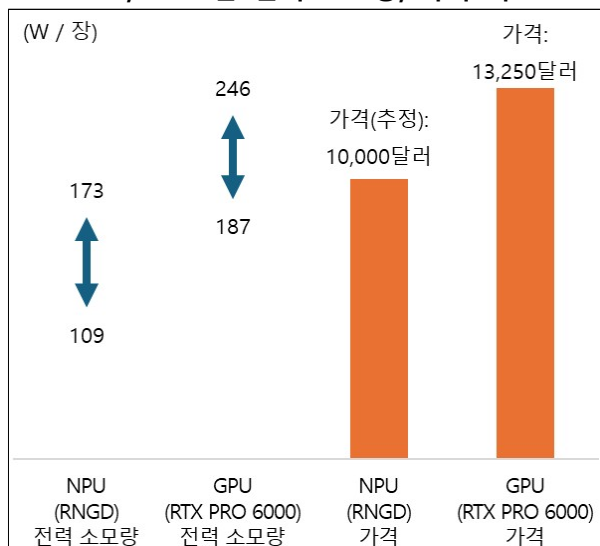
- 모델 개발사의 저가 전략이 수익 확대에 이어지기 위해서는 신규 수요 창출과 이용 증가에 따른 토큰 사용량 확대가 관건
 - 낮아진 가격으로 모델 개발사의 토큰당 이익이 감소하더라도, 토큰 사용량이 큰 폭으로 증가 시 전체 이익은 증가 가능
- 가격 경쟁 심화로 원가 경쟁력이 AI 모델 개발사의 승부처로 부상함에 따라 토큰 원가 절감이 수익성 확보의 핵심 과제
 - OpenAI와 Anthropic의 매출액 대비 추론 비용*이 60% 이상으로 추정되는 등 모델 개발 완료 후에도 인프라에 투입되는 원가 비중이 크므로 절감 필요
 - * 학습 비용과 별도로, 학습된 AI 모델이 이용자 질의에 대한 답변을 생성할 때 소요되는 추론 연산 자원 이용료, 전력비 등의 비용
 - 인프라 비용 절감뿐만 아니라 모델 효율 개선을 통한 토큰 원가 절감도 과제
- 추론 비용 절감을 위한 저전력·저비용 칩 수요가 확대될 경우, 국내 NPU* 기업에는 시장 선점 및 성장 기회가 될 전망
 - * NPU(Neural Processing Unit): 추론에 필요한 연산만 효율적으로 수행하도록 설계된 특화형 반도체로, 전력 효율성과 비용 측면에서 강점
 - NPU는 기존 GPU 대비 약 25% 저렴하고 전력 소모량도 약 30%~40% 적은 수준

‘25년 OpenAI(좌), Anthropic(우) 재무지표(추정)



자료 : Stanford HAI(2026.4), "Artificial Intelligence Index Report 2026" 및 언론보도 참고하여 산업은행 재구성

NPU, GPU 간 전력 소모량, 가격 비교



자료 : Lablup(2026.7), "RNGD meets Backend.AI", 산업은행 재구성