

DeepSeek發 혁신, AI 생태계 다원화 촉발

저자 : 김은수 이우근



푸른 고래를 로고로 하는 DeepSeek가 AI 생태계에 파장을 일으키고 있다. (DALL-E 생성 이미지)

- 모델 공개 이후 이어진 세 번의 추가 혁신 공개
- 딥시크 이후, 시장 참여자들은 포지셔닝 변화 모색
- AI 생태계는 소수 독점에서 다원화
- AI 협업 생태계를 최적화하며 고객가치를 모색할 시점

올해 초 중국 신생기업 DeepSeek(이하 딥시크)의 추론 모델 R1 공개는 AI 업계에 충격을 안겼다. 고성능 AI 개발에 수천억 원 규모의 방대한 컴퓨팅 자원이 필수라는 통념을 깨뜨린 이 사건은, 저비용으로도 고성능을 달성하는 대안 경로를 실증하고 오픈소스로 확산시켰다. 당시에는 AI 모델 기업 간 벤치마크 경쟁으로 여겨졌으나, 딥시크는 한 달 만에 저비용 API 구현 기술과 비용구조까지 공개해, 고가 API 기반 폐쇄적 빅테크들의 서비스 방식에 신선한 도전과 충격을 주었다. 딥시크로 인해 AI 경쟁은 ‘누가 더 뛰어난 모델을 만드느냐’와 함께 ‘누구의 방식이 AI 시장에 더 큰 영향력을 가질 것이냐’로 확장되고 있다.

모델 공개 이후 이어진 세 번의 추가 혁신 공개

딥시크 임팩트는 모델 공개의 단발성 이벤트로 끝나지 않았다. 추가로 세 차례에 걸친 핵심기술 공개를 통해 AI 시장의 판도를 전략적으로 재편하며 경쟁 구도를 자신들에게 유리한 방향으로 유도하고 있다.

우선, 1월 22일 딥시크는 R1 논문을 통해 추론모델 개발 방식을 최초로 공개했다. 오픈AI가 지난해 9월 GPT-o1 시리즈 출시와 함께 강화학습 기반 추론모델 시대를 선언했으나 구체적인 방법론은 비공개로 유지했던 상황에서, 딥시크는 유사한 수준의 추론 능력을 구현하는 레시피까지 공개했다. 마크 쉐인 오픈AI CRO(최고연구책임자)가 베일에 가려져 있던 GPT-o1의 핵심 개발 원리 중 일부를 딥시크가 독자적으로 개발했음을 인정하면서, 추론모델은 오픈AI 독점에서 다양한 주체가 참여 가능한 영역으로 확장될 것이 예고됐다.

두 번째로, 딥시크는 서비스에 ‘Deep Think’ 기능을 최초로 탑재했다. 이 기능은 AI가 프롬프트를 분석하고 지식을 활용해 결론에 도달하는 전체 사고 과정을 투명하게 공개한다. 오픈 AI가 사고과정을 비공개로 유지하고 STEM 영역의 복잡한 추론에만 제한적 활용을 권장했던 것과는 대조적이다. Reddit, X 등 소셜미디어에서는 ‘단순히 답을 제공하는 것을 넘어 AI의 사고 과정을 따라갈 수 있어 딥시크가 GPT-o1보다 더 유용하고 인간적’이라는 호평이 이어졌다. 이는 추론모델의 활용 범위를 창작, 기획 등 일상 업무로 확장했고, 경쟁사들의 유사 서비스 출시를 촉진했다.

세 번째로, 2월 24일부터 3월 1일까지 딥시크는 ‘오픈소스 워크’를 통해 또 다른 혁신을 선보였다. 저가 API 구현을 위한 핵심기술 다섯 가지를 이례적으로 엔지니어링 세부사항까지 상세히 공유한 것이다. 이는 고가의 최신 GPU 없이도 효율적인 AI 서비스가 가능함을 실증적으로 보여주었다. 딥시크는 이 기술을 통해 AI 학습비용을 95% 절감하고, 피크타임 기준 추론 서비스 마진을 545%까지 높였다고 발표했다. 하드웨어 최적화만으로도 AI 서비스의 수익성이 크게 개선 가능하다는 사실을 입증한 것이다.

이러한 세 단계의 전략적 기술 공개로 딥시크는 AI 경쟁을 ‘최고 성능 모델 개발’에서 ‘효율성과 서비스 최적화’ 쪽으로 다원화시켰다. 이제 AI 시장 참여자들은 기존 빅테크 주도의 고성능 접근법과 딥시크의 저비용 고효율 접근법에서 어느 방식에 전략을 맞출지, 두 방식의 장점을 어떻게 융합할지 고민하는 국면에 접어들었다.

사고 과정을 보여주는 <think> 기능

Find out two 3-digit palindromic numbers that added result in a 4-digit palindromic number.
Note: the first digit of any of the above mentioned three numbers can't be 0.



DeepSeek R1

<think> Okay, so I need to find two 3-digit palindromic numbers that, when added together, result in a 4-digit palindromic number. And none of the numbers can start with a zero. Let me break this down.

First, what's a palindromic number? It's a number that reads the same forwards and backwards. For example, 121 or 1331. So, a 3-digit palindrome would have the form ABA, where A and B are digits, and A can't be zero. Similarly, a 4-digit palindrome would have the form ABBA, right? Let me check: 1221 is a palindrome, yeah. So ABBA is

출처: 딥시크

딥시크 이후, 시장 참여자들은 포지셔닝 변화 모색

딥시크의 등장은 AI 산업 전반에 걸쳐 기업들의 전략 조정을 촉발하고 있다. AI 가치사슬의 주요 단계별로 기업들의 대응 방식을 살펴보자.

빅테크 기업들이 딥시크 이후 오픈소스 기반 협업 비중을 높이고 있다. 구글은 폐쇄형 모델인 제미니(Gemini) 외에 경량화된 오픈소스 모델인 젤마(Gemma)를 업그레이드하여 출시했으며, 메타는 개방형 모델을 유지하기 위해 딥시크의 오픈소스가 갖는 보안 위험을 방지하는 대책을 내놓기도 했다. 폐쇄형 전략의 대표주자 오픈AI도 일본·한국과의 협력을 확대하는 한편, 케빈 웨일 오픈AI CPO(최고제품책임자)는 개발자 생태계 확장을 위해 일부 비공개 모델의 오픈소스화를 검토 중이라고 밝혔다. 중국 빅테크 기업의 전략 선회는 더욱 극적이다. 리엔홍 바이두 CEO는 불과 몇 달 전 오픈소스는 일종의 돈 낭비라고 비하했지만, 딥시크 성공 이후 그들의 AI 모델인 원신(文心) 4.5의 오픈소스화를 전격 발표하며 ‘우수 모델의 오픈소스화가 응용을 크게 촉진한다’는 상반된 입장으로 급선회했다.

한편, 인프라 영역에서는 엔비디아 GPU 독점에 대한 도전이 본격화되고 있다. 가장 주목할 만한 움직임은 UALink(Ultra Accelerator Link) 연합의 등장이다. 아마존(AWS), MS, 메타, 구글이 주도하고 65개 이상 기업이 참여한 이 이니셔티브는 딥시크가 보여준 GPU 최적화에 탄력을 받아 더욱 강화되고 있다. UALink는 개방형 표준으로, 기존 엔비디아 GPU에 특화된 통신 방식(NVLink)에 대한 의존도를 낮출 수 있다. UALink에 참여한 기업들은 이를 통해 데이터센터 구축 비용과 유연성을 개선할 수 있는데, 올해초 표준이 공개된 후 내년부터 실제 AI 클러스터에 도입될 예정이다.

AI 모델 유형에서는 ‘더 큰 모델’에서 ‘더 효율적이고 특화된 모델’ 적용이 활발해지고 있다. 프랑스 미스트랄은 80개 이상의 프로그래밍 언어를 지원하는 코드 생성 특화 모델, 아랍어나 다양한 인도계 언어 특화 모델이나 간단한 대화로 이미지를 처리할 수 있는 멀티모달 특화 모델을 각각 제공한다. 중국에서는 딥시크와 같은 AI 스타트업을 포함해 산업별로 특화된 중소형 모델 기반 AI 상품이 300개 이상 사용 중이다. MS는 이런 변화에 대응해 ‘애저 AI 스튜디오’를 다중 모델 통합 플랫폼인 ‘애저 AI 파운드리’로 확장했으며, 플랫폼 워크로드의 30% 이상이 오픈AI 외 모델을 활용하고 있다. 딥시크의 모델 최적화 기술 공개로 각 기업들의 직접 모델 운영이 용이해지면서 MS 애저, 구글클라우드, AWS, 알리바바 클라우드 등은 API 제공에서 AI 최적화 및 다중 모델 통합 관리를 위한 특화 서비스로 수익 모델을 확장하고 있다.

응용 서비스에서는 ‘범용 모델 API 서비스 중심’에서 ‘특화 모델과 하드웨어-소프트웨어 통합 접근’이 시도되고 있다. 대부분 기기(PC, 휴대폰 등)의 AI 서비스는 대형 언어 모델을 탑재하는 수준에 머물러 있었다. 그러나 AI 모델이 기계장치의 작동에 맞게 개발되고 유기적으로 통합됐을 때 수행 가능한 작업 범위가 넓어지고 성능이 비약적으로 높아질 수 있다. 로봇 제조 스타트업 Figure AI는 AI와 로봇 통합을 위해 지난 2월 오픈AI와의 파트너십을 종료한 뒤 자체 AI 모델(Helix)을 발표했다. 외부 모델 소싱이 아닌 자체 모델을 통해 로봇이 물건을 쉽게 집어 올리거나 상체를 부드럽게 제어하고 로봇 간 작업 협업에 초점을 맞추는 등 AI를 로봇 성능 향상에 특화시키고 있다.

자체모델(Helix)이 탑재된 Figure 02의 시연 모습



자체모델(Helix)이 탑재된 두 로봇이 협업하여 식료품들을 냉장 실온 보관으로 분류 정리하고 있다.

출처: Figure AI

AI 생태계는 소수 독점에서 다원화

2020년대 초반까지 빅테크 기업들이 독점했던 AI 생태계의 권력 구조가 보다 입체적으로 변화하고 있다.

빅테크 기업들은 여전히 AI 기술 혁신과 대규모 투자에서 중요한 역할을 하고 있다. 오픈AI의 강력한 GPT 시리즈, 구글의 제미니, 메타의 라마 등은 계속해서 AI의 경계를 확장하고 있다. 그러나 이들의 위치는 독점적 지배자에서 생태계 내 핵심 참여자로 변화하고 있다. 동시에 오픈소스 커뮤니티는 다양한 특화 모델과 효율화 기술을 접목하며 새로운 성능의 혁신을 시도하고 있다. 이 두 세력 간의 관계는 단순 경쟁이 아닌, 다양한 상호작용을 보이고 있다. 빅테크 기업들이 일부 모델과 도구를 오픈소스화하는 한편, 오픈소스 커뮤니티에서 개발된 효율화 기술이 빅테크의 시스템에 통합되는 사례도 늘고 있다. 이는 빅테크와 오픈소스 커뮤니티 간의 경쟁과 협력이 공존하는 새로운 균형을 의미한다.

AI 산업은 소수 기업이 모든 것을 통제하는 구조에서, 다양한 참여자가 각자의 강점을 살려 공존하는 생태계로 진화하고 있다. 빅테크는 최첨단 연구와 대규모 서비스 제공에, 오픈소스 커뮤니티는 효율화와 특화 모델 개발에, 각 산업의 전문가들은 도메인 지식을 AI와 결합하는 데 집중하는 모습이다. 딥시크가 촉발한 생태계 다원화는 AI 기술의 민주화를 가속하고 있으며, 다양하고 혁신적인 AI 응용으로 이어질 전망이다.

딥시크가 촉발한 AI 생태계 변화

구분		(과거) 빅테크 주도 폐쇄형 생태계	(현재) 딥시크 등장으로 추가된 개방형 생태계
시장 변화	경쟁 방식	'더 크고, 더 강력한 모델' 개발	'더 효율적이고, 더 특화된' 개발
	주요 참여자	소수 빅테크(오픈AI, 구글, MS, 메타, 아마존 등)	빅테크 + 다양한 스타트업 + 오픈소스 커뮤니티
가치사슬 변화	데이터 수집	빅테크의 독점적 데이터 소유	합성 데이터와 증류 기법으로 데이터 격차 감소
	모델 개발	수천억 원의 투자 필요, 폐쇄적 개발	오픈소스 레시피로 비용 대폭 절감, 협업적 개발
	핵심 인프라	엔비디아 GPU 92% 점유, 수직적 통합	UALink 등 대안 생태계, 이종 GPU 활용
	응용	범용 대형 모델 중심, API 서비스	특화 모델 다양화, 하드웨어-소프트웨어 통합

AI 협업 생태계를 최적화하며 고객가치를 모색할 시점

AI 생태계가 다원화되면서 서비스 소비 주체에 머물던 기업들에게 내부혁신과 특화 서비스 참여 기회가 열리고 있다. 개별 기업은 AI 생태계 변화 환경에서 다음을 유의해야 한다.

첫째, 자원 활용의 유연성을 높여야 한다. 기업은 자원 수준에 맞춰 ▲자체 개발 ▲오픈소스 모델의 미세 조정 ▲API 활용 중 최적안을 선택해야 한다. 자체 개발로 외부의존도를 낮추거나, 오픈소스 활용으로 절감한 초기 비용을 AI 친화적 데이터 구축과 추론 효율성 향상에 재투자하거나, API를 통해 신속하게 서비스를 적용하는 등 각사의 상황에 맞는 접근이 필요하다. 더불어 UALink와 같은 개방형 인프라도 주목할 가치가 있다. 단위 서버와 GPU 비용을 절감해 더 많은 계산 자원을 확보하거나, 특정 기능에 맞는 유연한 인프라 구성이 가능하기 때문이다.

둘째, 경쟁우위의 원천이 모델에서 도메인 지식으로 전환될 가능성에 대비해야 한다. 딥시크의 혁신으로 AI 모델은 차별화 요소가 아닌 필수적 기반(Baseline)으로 자리잡고 있다. 진정한 경쟁 우위는 특정 산업과 고객에 대한 깊은 이해에서 비롯될 수 있다. 산업 내 암묵지를 AI 모델로 신속히 구축하는 기업은 각 산업에서 선도적 위치를 차지할 수 있다. 예를 들어, 산업 생산에서 공정 데이터로 품질 문제와 장비 고장을 예측하는 특화 AI 모델은 저성장 시대의 핵심 자산이 될 수 있다.

셋째, AI의 진정한 경쟁력은 비용 절감이 아닌 서비스 가치에 있다. AI 서비스 가격은 확산을 가로막는 분명한 장벽이다. 그러나, 스마트폰이 고가에도 피쳐폰을 대체했듯 AI도 가격보다 제공 가치가 결정적일 수 있다. 빅테크 기업들이 딥시크 이후에도 대규모 투자를 지속하는 것은 기존 범용 AI가 다수 고객이 적극 수용할 만한 수준에는 이르지 못했다는 판단 때문이다. 기업 고유의 AI 서비스 가치를 높이는 것이 원가 절감보다 중요하다.

AI 개발과 응용의 블랙박스가 서서히 열리고 있다. 이제 AI 혁신은 빅테크 기업의 전유물이 아닌, 생태계에 속한 모든 기업의 몫이다. 더 이상 수동적 관찰자로 머물기보다 AI 생태계 변화 속에서 개별 기업이 만들 수 있는 가치를 발견하고 구체화할 시점이다.

딥시크 혁신의 뒷 이야기

반도체 수출 제재가 촉발한 AI 터닝포인트

딥시크는 표면상 2023년 설립된 기업이지만, 10년 이상의 준비와 전략적 진화가 있었다. 주목할 점은 량원평 CEO가 GPU 시장 과열 이전부터 대규모 컴퓨팅 인프라에 적극 투자했다는 사실이다. 당시 그도 고사양 중심 AI 개발의 주요 참여자였다. 그러나 2022년 10월 미국의 대중국 반도체 수출 제재로 상황이 급변했다. 고사양 인프라 확보가 불가능해지자, 량원평은 이 위기를 혁신의 기회로 전환했다. 기존의 규모 확장 전략을 버리고, 선도 기업들이 이미 개발한 기술과 지식을 최대한 활용하는 동시에 저비용 고효율 개발에 초점을 맞추는 오픈소스 전략으로 선회했다. 이러한 환경적 제약이 역설적으로 딥시크의 혁신을 이끌었고, 글로벌 AI 생태계에 새로운 가능성을 열었다.

알고리즘 혁신보다 시스템 통합 최적화에 집중

대부분 AI 기업이 알고리즘 개발에 집중할 때, 딥시크는 하드웨어 튜닝과 시스템 최적화에 특화했다. 모회사 환방은 중국 4대 쿼트 기업으로, 고빈도 거래 경험에서 얻은 연산효율화 및 통신 병목 최소화 기술을 AI 개발에 접목했다. 이들은 일찍부터 표준통신 방식을 우회해 GPU 자원을 최대로 활용하였으며, 자체 개발이 익숙한 중국 AI 인력의 강점도 활용했다.

단기 수익보다 생태계 주도권 확보를 위한 조직 운영

딥시크는 모회사의 100% 출자로 운영돼 투자 수익에 대한 외부 압박 없이 기초 연구에 집중했다. 창업자 량원평은 ‘경력이 없는 사람이 반복 탐색과 고민을 통해 현 상황에 맞는 해결책을 모색한다’는 인재관을 바탕으로 20-30대 젊은 인재 중심의 R&D팀을 구성하고, KPI를 사전 설정하지 않는 느슨한 관리를 장려했다. 그 결과 1~2년차 중심의 상향식 개발로 MLA, GRPO 등 혁신이 가능했다.