DeepSeek 연구 결과 종합

1. DeepSeek 일반 정보

DeepSeek(중국어 간체자: 深度求索, 병음: Shēndù Qiúsuǒ)는 중국의 인공지능 연구 기업이자 회사의 제품명입니다. 2023년 5월에 설립되었으며, 중국의 헤지펀드인 High-Flyer의 대규모 자금 지원을 받고 있습니다. 두 회사 모두 량원펑(Liang Wenfeng)이 설립하고 운영하고 있으며, 저장성 항저우에 본사를 두고 있습니다.

설립 배경

- 2015년: 량원펑이 동문 2명과 함께 중국 최대 퀀트 헤지펀드 하이플라이어(High-Flyer) 설립
- 2019년: AI 기반 알고리즘 트레이딩으로 자산운용규모(AUM) 100억 달러(약 15조 원) 돌파
- 2021년: 10,000개의 NVIDIA H100 GPU 클러스터 구축해 대규모 AI 실험 기반 마련
- 2023년 5월: 금융 모델의 한계를 넘어 범용 AI 기술 개발을 위해 딥시크(DeepSeek) 연구실을 독립법인으로 분사

주요 모델 출시 타임라인

- 2023년 11월: 코딩 특화 모델 'DeepSeek Coder' 출시
- 2023년 11월: 범용 대규모 언어모델 'DeepSeek LLM' 시리즈 공개 (7B 및 67B 파라미터 모델)
- 2024년 4월: 수학 문제 해결에 특화된 'DeepSeek Math' 7B 모델 공개
- 2024년 5월: 성능 향상 및 비용 절감에 초점을 맞춘 'DeepSeek-V2' 시리즈 출시
- 2024년 12월: 오픈소스 비전-언어 모델 'DeepSeek VL' 출시
- 2024년 12월: 671억 파라미터 규모의 'DeepSeek-V3' 모델 출시
- 2025년 1월 20일: 오픈소스 추론 모델 'DeepSeek R1' 공개

인력 구성

- DeepSeek의 연구·개발(R&D) 인력은 139명에 불과함
- 이는 챗GPT 개발사 오픈AI의 연구원 1200명, 마이크로소프트 코파일럿 개발사 마이크로소프트의 7000명, 제미나이 개발사 구글의 5000명과 비교됨
- 딥시크 창업자 량원펑을 비롯한 중국인 연구자·엔지니어 150명과 데이터 자동화 연구팀 31명이 개발에 참여

2. DeepSeek-R1 훈련 방법

DeepSeek-R1은 혁신적인 훈련 방법을 통해 적은 자원으로도 높은 성능을 달성했습니다. 주요 훈련 방법은 다음과 같습니다:

MoE(Mixture of Experts) 아키텍처

- DeepSeek-R1은 MoE 아키텍처를 사용하여 모델 효율성을 크게 향상시킴
- MoE는 특정 작업에 필요한 파라미터만을 선택적으로 활성화함으로써 계산 효율성을 높임
- V3는 MoE를 통해 총 6710억개 파라미터 중에서 약 5%만 사용
- 이를 통해 모델은 답변 전문성을 높이면서도 수많은 파라미터를 모두 활성화할 필요가 없어짐

지식 증류(Knowledge Distillation)

- 기존에 개발했던 뛰어난 모델(DeepSeek-R1)로, 새로운 모델인 DeepSeek-V3를 개발하는 지식 증류 기법 활용
- 서로 다른 모델끼리 지식을 전하는 기법으로, 주로 큰 모델에서 작은 모델로 지식을 전달하는 프로세스
- 이를 통해 추론 능력을 그대로 활용하면서도 모델 크기와 연산량을 줄임

하이브리드 학습 접근법

- 지도학습 미세조정(SFT, Supervised Fine-Tuning)과 강화학습(RL, Reinforcement Learning)을 하이브리드로 결합
- SFT는 데이터 셋에 모델을 튜닝하는 방식으로, 비교적 예측 가능한 방식으로 모델 성능을 향상
- 강화학습은 사용자의 피드백을 통해 모델을 점진적으로 개선하는 접근법
- 이 두 접근법을 결합하여 기존 모델들이 가지고 있던 한계를 넘어서는 모델을 개발

멀티토큰(Multi-Token) 기법

- 보통 AI 모델은 문장을 조각(Token)으로 나눠 읽는 반면 R1은 문장 전체를 하나로 처리
- 이를 통해 생성속도가 2배 더 빠르고 답변 정확도는 90%로 매우 높음
- 문장 전체를 컨텍스트로 이해하기 때문에 더 정확한 추론이 가능

그룹 상대 정책 최적화

- R1이 채택한 강화학습은 '절대적 평가 모델' 대신 '그룹 점수'라는 새로운 알고리즘을 적용
- 전자는 어떤 데이터를 학습해야 할지 지정해주는 방식
- 후자는 여러 행동을 그룹으로 묶어 비교하고 가장 좋은 결과를 수렴해서 찾아내주는 방식

3. DeepSeek의 비용 효율성 및 빅테크 대비 저비용 달성 정도

DeepSeek는 혁신적인 기술과 접근법을 통해 기존 빅테크 기업들보다 훨씬 저렴한 비용으로 AI 모델을 개발하고 서비스를 제공하고 있습니다.

개발 비용

- DeepSeek-V3 모델 개발에 558만 달러(약 84억원)의 비용이 소요됨
- 이는 OpenAI의 GPT-4 개발 비용으로 알려진 1억 달러와 비교하면 약 18배 저렴한 수준
- 오픈AI가 사전 훈련 모델을 1~2년 주기로 개발하는 것에 비해 훨씬 빠른 속도로 개발

하드웨어 활용

- 개발 과정에서 고가의 GPU 대신 H800 그래픽처리장치를 사용
- 성능이 H100 절반에도 미치지 못하며, 가격도 훨씬 저렴한 H800 GPU만으로 DeepSeek-V3 개발을 완수
- 이는 고가 GPU를 감당할 수 있는 기업만이 아니라 중소기업과 비(非) 미국 기업도 경쟁력 있는 AI 모델을 개발할 수 있다는 가능성을 보여줌

API 이용 비용

- DeepSeek API는 입력 토큰 100만 개당 $0.55, 출력 토큰 100만 개당 $2.19라는 파격적인 가격 제시
- 이는 오픈AI의 O1모델 API 이용료 입력 토큰 100만 개당 $15, 출력 토큰 100만 개당 $60 대비 약 1/30 수준의 가격
- 실사용 테스트에서 OpenAI 모델 비용의 1/20 수준으로 서비스 제공

추론 비용

- DeepSeek-R1 모델의 추론 비용은 OpenAI o1 모델 대비 90~95% 정도 절감
- 혼합전문가(Mixture of Experts, MoE) 구조와 강화학습 기반 훈련 방식을 도입하여 모델의 연산 비용을 90%까지 절감
- 이를 통해 기업들의 AI 도입 비용을 획기적으로 낮출 수 있는 수준 제공

비용 절감 요인

- 금융 데이터 분석에서 축적한 알고리즘 최적화 기술을 접목해 하드웨어 의존도를 60% 이상 낮추는 혁신 달성
- MoE 구조는 특정 작업에 필요한 파라미터만을 선택적으로 활성화함으로써 계산 효율성을 크게 높임
- 강화학습 기반 훈련은 모델의 성능을 최적화하는 데 필요한 데이터와 시간을 줄여줌
- 이러한 기술적 혁신은 DeepSeek AI가 고성능 모델을 유지하면서도 운영 비용을 대폭 낮출 수 있게 함

4. DeepSeek의 기술 공개 여부

DeepSeek는 오픈소스 전략을 적극적으로 채택하여 대부분의 모델과 기술을 공개하고 있습니다.

오픈소스 라이선스

- DeepSeek는 대부분의 모델을 MIT 라이선스로 공개
- MIT 라이선스는 상업적 용도로도 제한 없이 사용할 수 있는 매우 자유로운 라이선스
- 전 세계 개발자들이 자유롭게 수정하고 상용화할 수 있도록 허용

공개된 모델

- DeepSeek Coder: 2023년 11월 2일에 연구원과 상업 사용자 모두에게 무료로 제공되는 첫 번째 모델
- DeepSeek LLM: 모델 코드는 MIT 라이선스에 따라 오픈 소스로 공개
- DeepSeek-R1: MIT 라이선스 하에 완전 공개되어 상업적 용도로도 제한 없이 사용 가능
- DeepSeek-V3: 모델 아키텍처와 훈련 방법에 대한 상세 정보 공개

오픈소스 생태계 기여

- 깃허브 기여자 수가 3개월 만에 2만 명을 돌파하는 놀라운 성장세
- 허깅페이스에 따르면, 2025년 1월 말 기준 딥시크의 R1 모델을 기반으로 617개의 새로운 모델이 생성
- 2월 기준 총 다운로드 수는 300만 건을 넘음
- DeepSeek의 오픈소스 정책은 단순히 코드 공개를 넘어 글로벌 협업 생태계를 구축하는 데 중점

오픈소스 전략의 의의

- AI 기술의 민주화를 가속화하고, 더 많은 기업과 개발자들이 첨단 AI 기술을 활용할 수 있는 기회 제공

- 글로벌 AI 생태계의 다양성과 발전을 촉진
- 특히 중국과 아시아 지역의 개발자들의 참여가 크게 증가

5. DeepSeek가 LLM 분야에 미친 영향

DeepSeek의 등장은 LLM 분야에 여러 측면에서 큰 영향을 미치고 있습니다.

가격 경쟁 촉발

- DeepSeek-V2 모델은 1백만 토큰당 1위안의 낮은 비용으로 운영될 수 있어 "AI계의 핀둬둬"라는 별칭을 얻음
- 바이트댄스, 텐센트, 바이두, 알리바바와 같은 다른 주요 기술 대기업들이 이 회사와 경쟁하기 위해 자사 AI 모델의 가격을 낮추기 시작
- 1월 말에는 엔비디아를 비롯한 빅테크 기업들의 주가 급락으로 이어짐
- 딥시크의 등장은 텐센트, 알리바바와 같은 중국 빅테크 기업 간의 AI 가격 경쟁에 불을 지핌

오픈소스 AI 생태계 활성화

- 오픈소스 AI 모델의 새로운 지평을 열었다는 평가
- 허깅페이스에 R1 모델을 기반으로 617개의 새로운 모델이 생성되고 300만 건 이상의 다운로드 기록
- 폐쇄형 AI와 오픈소스 AI의 장단점에 대한 더 큰 논의의 장 마련
- 기업들이 오픈AI 같은 거대 폐쇄형 AI보다 필요에 맞게 조정할 수 있는 소형 AI를 선호하는 추세 형성

저비용 고효율 AI 연구 촉진

- 저비용 고효율 AI 연구에 대한 관심을 촉발하면서, 주요 기업들의 '가성비' 모델 개발을 이끌어내는 촉매제 역할
- 네이버는 하이퍼클로바X의 신규 버전을 공개하며 기존 대비 40% 수준의 작은 크기로도 더 뛰어난 성능을 발휘하고 운영 비용이 50% 이상 절감됨을 강조
- 고성능 GPU를 감당할 수 있는 기업만이 아니라 중소기업과 비(非) 미국 기업도 경쟁력 있는 AI 모델을 개발할 수 있다는 가능성 제시

AI 기술 접근성 향상

- AI 기술의 민주화를 가속화하고, 더 많은 기업과 개발자들이 첨단 AI 기술을 활용할 수 있는 기회 제공
- 스타트업과 중소기업들의 AI 혁신을 가속화
- 중국 외 지역의 개발자들에게도 고성능 AI 모델에 대한 접근성 향상

빅테크 기업들의 대응 변화

- 오픈AI의 샘 알트만 CEO도 "새로운 경쟁자가 등장한 것은 분명히 신선한 자극이다"라고 언급
- 애플의 팀 쿡 CEO는 "일반적으로 효율을 주도하는 혁신은 긍정적이며, 딥시크의 모델에서도 그런 점을 확인할 수 있다"고 설명
- 마이크로소프트의 사티아 나델라 CEO는 "AI가 더 효율적이고 접근 가능해질수록 사용량이 기하급수적으로 증가할 것이며, 이는 마이크로소프트와 같은 하이퍼 스케일러들에게 기회가 될 것"이라고 강조
- 기존 빅테크 기업들이 더 개방적인 정책을 채택할 가능성 증가

도전과 한계

- 보안 취약성 문제: 사용자 및 기기 정보를 암호화하지 않은 채 전송해 중간자 공격이나 스니핑과 같은 해킹에 취약
- 개인정보 보호 체계 부실: 개인정보를 동의 없이 중국 기업에 전송한 사실이 확인되면서 파장이 커짐
- 윤리적 안전장치 미비: AI 모델이 생물무기 제조법, 악성코드가 삽입된 피싱 이메일, 반유대주의 성향의 히틀러 옹호 선언문까지 생성하는 것으로 확인됨
- 이러한 문제로 인해 여러 국가에서 접근 제한 조치가 취해지고 있음