# DeepSeek 보고서: LLM 분야의 혁신과 비용 효율성

## 서론: LLM 환경에서 DeepSeek의 등장

2023년에 설립된 중국의 AI 기업 DeepSeek은 경쟁력 있는 비용 효율적인 AI 솔루션을 제공하여 고급 언어 모델에 대한 접근성을 높이는 것을 목표로 빠르게 성장하고 있습니다1. DeepSeek은 효율성과 성능 최적화에 중점을 두어 시장에서 차별화를 꾀하고 있으며2, OpenAI와 같은 주요 기업의 모델과는 대조적으로 오픈 소스 방식을 채택하고 있습니다3. 본 보고서는 DeepSeek의 비용 효율성, 핵심 기술 혁신, 기술 공개 범위, 그리고 LLM 분야의 언어 생성, 추론 효율성, 모델 경량화에 미치는 영향에 대해 분석합니다.

DeepSeek이 미국 거대 기술 기업들에 도전하는 중국 AI 기업으로 부상한 것은4 글로벌 AI 환경의 잠재적인 변화를 시사하며, 이는 다양한 지역으로부터의 경쟁과 혁신 증가를 의미합니다. 또한, DeepSeek의 모델 출시 시점이 OpenAI와 같은 주요 경쟁사들의 발표와 유사하거나 직후인 경우가 많다는 점은3 LLM 분야의 치열한 경쟁 환경과 빠른 혁신 속도를 보여줍니다. 이러한 상황은 AI 기술 전반의 발전을 가속화할 수 있습니다.

## 빅테크 대비 DeepSeek의 비용 효율성

### 토큰당 가격 분석

DeepSeek R1과 OpenAI 모델(o1, o3-mini, GPT-4)의 토큰당 입력 및 출력 비용을 비교하면 상당한 차이를 확인할 수 있습니다. DeepSeek R1의 입력 비용은 100만 토큰당 캐시된 경우 0.14달러, 캐시되지 않은 경우 0.55달러이며, 출력 비용은 100만 토큰당 2.19달러입니다1. 반면 OpenAI o1의 입력 비용은 캐시된 경우 7.50달러, 캐시되지 않은 경우 15.00달러이며, 출력 비용은 100만 토큰당 60.00달러로 나타났습니다1. 이는 DeepSeek R1이 OpenAI o1 비용의 약 5% 수준으로 운영될 수 있음을 시사합니다1.

OpenAI o3-mini의 경우 100만 토큰당 4.40달러로3, o1보다는 비용 효율적이지만 여전히 DeepSeek R1보다는 비쌉니다. Gemini 2 Flash는 100만 출력 토큰당 0.7달러로 가장 저렴한 옵션 중 하나입니다3. 다만, DeepSeek R1이 o3-mini보다 토큰당 비용은 저렴하지만(2.19달러 vs 4.40달러), 추론 과정에서 훨씬 더 많은 토큰(최대 3배)을 사용할 수 있어 특정 작업에서는 OpenAI 모델이 실제로 더 비용 효율적일 수 있다는 점을 고려해야 합니다10. 이는 토큰당 비용뿐만 아니라 실제 토큰 사용량을 함께 고려해야 LLM의 총 비용 효율성을 정확하게 평가할 수 있음을 보여줍니다.

### 전체 학습 비용 비교

DeepSeek은 자사의 모델(V3 등)을 약 560만 달러의 비용으로 2,000개의 Nvidia H800 GPU를 사용하여 학습했다고 주장합니다7. 이는 GPT-4의 8천만~1억 달러, Meta LLaMA 3의 16,000개 H100 GPU에 필요한 비용과 비교했을 때 상당히 낮은 수치입니다7. 다만, 560만 달러라는 비용이 전체 개발 비용이 아닌 최종 학습 실행 비용만을 의미할 수 있다는 추측도 있습니다13. DeepSeek은 DeepSeek-R1의 경우 OpenAI GPT-4의 1억 달러에 비해 약 1,500만 달러의 예산으로 유사한 성과를 달성했다고 주장합니다15.

### 표: 토큰당 비용 비교 (100만 토큰 기준)

| **모델** | **입력 비용 (캐시)** | **입력 비용 (미캐시)** | **출력 비용** |
| --- | --- | --- | --- |
| DeepSeek R1 | $0.14 | $0.55 | $2.19 |
| OpenAI o1 | $7.50 | $15.00 | $60.00 |
| OpenAI o3-mini | - | - | $4.40 |

DeepSeek의 현저히 낮은 토큰당 가격은 비용에 민감한 사용자 및 기업을 유치하려는 전략적 초점을 시사하며, 이는 기존 업체의 가격 모델에 큰 영향을 미칠 수 있습니다. DeepSeek R1의 출력 비용이 OpenAI o1의 60달러에 비해 2.19달러라는 점은 DeepSeek이 훨씬 저렴한 대안을 제공하려는 의도적인 노력을 보여줍니다. 이는 이전에는 비용 때문에 AI 도입이 어려웠던 애플리케이션 분야에서 AI 채택을 증가시킬 수 있습니다.

하지만 토큰당 비용과 실제 비용 효율성 간의 차이10에서 언급된 토큰 사용량 차이로 인한)는 LLM 비용 평가의 복잡성과 사용자의 특정 사용 사례를 고려해야 할 필요성을 강조합니다. DeepSeek의 토큰당 요금이 낮더라도 동일한 결과를 얻기 위해 훨씬 더 많은 토큰이 필요한 경우, 전체 비용 절감 효과는 처음에 예상했던 것보다 적을 수 있습니다. 이는 다양한 모델과 작업에 대한 토큰 효율성을 벤치마킹하고 분석하는 것이 중요하다는 점을 시사합니다.

잠재적인 과소평가 가능성에도 불구하고 DeepSeek 모델의 현저히 낮은 보고된 학습 비용은 혁신적인 학습 기술과 효율적인 리소스 활용을 나타냅니다. 이는 새로운 플레이어의 LLM 개발 진입 장벽을 낮춤으로써 LLM 개발의 미래에 영향을 미칠 수 있습니다. DeepSeek이 GPT-4와 같은 모델과 유사한 성능을 훨씬 적은 학습 비용으로 달성할 수 있다면, 이는 학습 방법론 또는 하드웨어 최적화에 획기적인 발전이 있음을 의미합니다. 이는 더 광범위한 조직과 연구자들이 고급 AI 개발에 더 쉽게 접근할 수 있도록 할 수 있습니다.

## 저비용 고성능 추론 모델(DeepSeek-R1) 훈련 방법

### R1-Zero 및 강화 학습(RL)

DeepSeek-R1은 기존 LLM 학습 방식과는 다르게 지도 학습(SFT)을 사전 단계로 거치지 않고 대규모 강화 학습(RL)만을 사용하여 초기 훈련을 진행하는 독특한 "R1-Zero" 방식을 사용했습니다4. R1-Zero는 보상 시스템에 기반한 시행착오 학습을 통해 순수하게 RL만으로도 고급 추론 능력을 개발할 수 있다는 가능성을 입증했습니다4. 추론 능력 향상을 위해 GRPO(Group Relative Policy Optimization)와 같은 특정 RL 알고리즘이 사용되었습니다21. 하지만 R1-Zero는 초기 단계에서 가독성 저하 및 영어와 중국어 혼용과 같은 문제점을 보였습니다16.

### DeepSeek-R1의 다단계 훈련 파이프라인

DeepSeek-R1은 R1-Zero를 기반으로 초기 RL 단계 *이후*에 소량의 콜드 스타트 데이터와 지도 학습(SFT)을 통합한 다단계 훈련 파이프라인을 도입하여 개발되었습니다16. 이러한 하이브리드 방식(RL + SFT)은 추론 성능을 더욱 향상시키고 R1-Zero에 비해 더욱 정제된 출력을 제공했습니다16. R1 훈련에는 긴 CoT(Chain of Thought) 예시와 인간 선호도 정렬이 포함된 선별된 데이터 세트가 사용되었습니다16.

R1-Zero가 순수하게 RL만으로 추론 능력을 개발하는 데 성공했다는 사실은 고급 LLM에 지도 학습이 필수적인 첫 번째 단계라는 기존의 통념에 도전합니다. DeepSeek이 초기 훈련 단계에서 대규모의 인간 레이블링 데이터에 의존하지 않고 강력한 추론 능력을 달성할 수 있었다는 점은 RL이 특정 AI 기능을 개발하는 데 있어 강력하고 잠재적으로 더 확장 가능한 접근 방식이 될 수 있음을 시사합니다. 이는 미래의 LLM 학습 방식에 변화를 가져올 수 있습니다.

R1-Zero에서 DeepSeek-R1으로 발전하면서 SFT를 추가한 것은 순수 RL이 강력한 추론 능력을 달성할 수 있지만, 대상화된 지도 학습을 결합하면 출력 품질 및 사용성과 같은 문제를 해결할 수 있음을 보여줍니다. 이러한 하이브리드 접근 방식은 RL의 새로운 기능 창출과 SFT의 인간 선호도에 맞춘 미세 조정 및 정렬이라는 서로 다른 학습 방법론의 강점을 활용하여 상호 보완할 수 있음을 입증합니다.

DeepSeek 훈련의 핵심 전략인 CoT(Chain of Thought) 추론에 대한 집중은1 특히 복잡한 작업에서 LLM 출력의 투명성과 설명 가능성을 향상시키는 데 중요한 역할을 합니다. 모델이 단계별 추론 과정을 생성하도록 훈련함으로써 DeepSeek은 AI의 의사 결정 과정을 더욱 이해하기 쉽고 신뢰할 수 있도록 만드는 것을 목표로 합니다. 이는 민감한 분야에서의 애플리케이션에 매우 중요합니다.

## 아키텍처 효율성: MoE 방식 및 그 이상

### MoE(Mixture-of-Experts) 아키텍처

DeepSeek 모델(R1 및 V3 포함)은 MoE(Mixture-of-Experts) 아키텍처를 활용합니다7. MoE는 많은 수의 "전문가" 하위 모델을 가지고 있으며, 게이팅 메커니즘을 통해 각 입력에 대해 이러한 전문가의 하위 집합(예: DeepSeek-R1 및 V3의 경우 6,710억 개 중 370억 개 파라미터)만 동적으로 선택하고 활성화하는 방식으로 작동합니다7. MoE의 장점은 추론 과정에서 컴퓨팅 비용을 크게 줄이고 처리 속도를 높이며, 컴퓨팅 요구 사항을 비례적으로 증가시키지 않고 모델 용량을 확장할 수 있다는 것입니다7. DeepSeek의 특정 MoE 구현인 DeepSeekMoE는 전문화를 개선하고 중복성을 줄이기 위해 세분화된 전문가 분할 및 공유 전문가 격리를 포함합니다31.

### 기타 아키텍처 혁신

MLA(Multi-Head Latent Attention)는 추론 중 KV(Key-Value) 캐시를 잠재 벡터로 압축하여 메모리 사용량을 줄임으로써 효율성에 기여합니다7. 또한 MTP(Multi-Token Prediction)는 여러 토큰을 동시에 예측하여 일관성 있는 응답 생성을 가속화하고 추론 속도를 높이는 역할을 합니다33.

### 하드웨어 및 시스템 수준 최적화

DeepSeek은 최첨단 칩 대신 잠재적으로 Nvidia H800 GPU와 같은 기존 하드웨어 최적화에 중점을 둡니다44. GPU에서 동시 연산 및 통신 처리를 위한 DualPipe 알고리즘을 개발했으며7, 보다 효율적인 GPU 명령어 실행을 위해 표준 CUDA 대신 PTX 프로그래밍을 사용합니다7. 또한 FP8 혼합 정밀도 연산을 사용하여 계산 비용과 메모리 사용량을 줄입니다7.

DeepSeek이 MoE 아키텍처에 크게 의존하는 것은 희소 활성화가 컴퓨팅 자원을 줄이면서 높은 성능을 달성하는 데 효과적이라는 전략적 판단을 시사합니다. DeepSeek 모델 전반에 걸쳐 MoE를 일관되게 사용하는 것은 이 아키텍처가 비용 효율성의 핵심이라고 믿기 때문입니다. 파라미터의 작은 부분만 활성화할 수 있기 때문에 빽빽한 모델의 전체 컴퓨팅 비용 없이 더 큰 용량의 모델을 만들 수 있습니다.

아키텍처 혁신(MoE, MLA, MTP)과 하드웨어/시스템 수준 최적화의 결합은 기술 스택의 여러 수준에서 효율성을 달성하기 위한 전체적인 접근 방식을 보여줍니다. DeepSeek은 단일 기술에만 의존하는 것이 아니라 모델 아키텍처, 메모리 관리, 예측 속도 및 하드웨어 활용을 최적화하여 상당한 비용 절감 및 성능 향상을 가져오는 시너지 효과를 창출합니다.

잠재적으로 덜 발전된 하드웨어(H800 GPU 등)로도 인상적인 결과를 달성할 수 있다는 DeepSeek의 능력은 특히 수출 규제에 직면했을 때 하드웨어 제약을 극복하는 데 있어 알고리즘 및 소프트웨어 최적화의 중요성을 강조합니다. 이는 AI 혁신이 최신 하드웨어에 대한 접근성뿐만 아니라 기존 자원의 영리한 엔지니어링 및 최적화에 의해 주도될 수 있음을 시사합니다. 이는 첨단 기술 확보에 제약이 있는 기업에 특히 관련이 있습니다.

## DeepSeek 기술의 공개 및 접근성

### 오픈 소스 모델 출시

DeepSeek은 DeepSeek-R1-Zero, DeepSeek-R1 및 DeepSeek-V3를 포함한 여러 모델을 오픈 소스로 공개했습니다3. 모델 가중치는 Hugging Face 및 GitHub와 같은 플랫폼에서 사용할 수 있습니다12. Llama 및 Qwen 아키텍처를 기반으로 한 15억, 70억, 80억, 140억, 320억, 700억 파라미터 모델과 같은 여러 파생 모델도 오픈 소스로 공개되었습니다4.

### 라이선스 조건

DeepSeek-R1 코드 저장소 및 모델 가중치는 MIT 라이선스 하에 라이선스가 부여되어 상업적 사용, 수정 및 파생 작업(파생 모델 학습 포함)이 허용됩니다25. 파생 모델의 경우 Qwen(Apache 2.0) 및 Llama(llama3.1 및 llama3.3)에서 파생되었으며 원래 라이선스를 유지합니다25.

### API 접근

DeepSeek API는 OpenAI API 형식과 호환됩니다50. API를 통해 접근 가능한 모델은 deepseek-chat(DeepSeek-V3로 업그레이드됨) 및 deepseek-reasoner(DeepSeek-R1)입니다50. 서버 리소스 제한으로 인해 API 계정에 새 자금을 추가하는 기능은 일시적으로 중단되었습니다51. DeepSeek-R1은 Amazon Bedrock에서 완전 관리형 서버리스 모델로도 제공됩니다47.

### 제한 사항 및 제약 사항

DeepSeek은 모델 학습 데이터 세트를 공개하지 않았습니다13. DeepSeek 모바일 앱과 관련된 보안 문제 및 중국으로의 데이터 전송 가능성에 대한 우려가 제기되었습니다13. 사용자 데이터가 중국 서버에 저장될 수 있다는 개인 정보 보호 정책도 언급되었습니다13.

특히 관대한 MIT 라이선스 하에 모델을 오픈 소스로 공개하려는 DeepSeek의 노력은 더 넓은 접근성, 수정 및 활용을 허용함으로써 AI 커뮤니티 내 협업 환경을 조성하고 혁신을 가속화합니다. OpenAI 에코시스템에 이미 익숙한 개발자에게 API 호환성은 진입 장벽을 낮추어 DeepSeek 모델을 더 쉽게 채택하고 실험할 수 있도록 합니다. 하지만 DeepSeek이 개방성을 장려하는 동시에 학습 데이터에 대한 투명성 부족 및 제기된 보안/개인 정보 보호 문제는 사용자 신뢰와 민감한 애플리케이션에서의 기술 채택에 영향을 미칠 수 있습니다.

## LLM 분야에 미치는 영향

### 언어 생성에 대한 영향

DeepSeek-R1 및 V3는 MMLU와 같은 언어 이해 작업에서 OpenAI o1, GPT-4o 및 Claude 3.5 Sonnet과 같은 모델과 경쟁력 있는 성능을 보여줍니다12. DeepSeek은 추론 작업에서 강력한 성능을 보이며 일관성 있고 맥락에 맞는 응답을 생성할 수 있습니다3. 특히 중국어와 영어에서 다국어 기능을 제공합니다17.

### 추론 효율성 혁신

DeepSeek은 OpenAI 모델에 비해 추론 비용이 훨씬 저렴합니다1. MoE, MLA, MTP와 같은 아키텍처 혁신은 더 빠른 추론 속도와 감소된 리소스 소비에 기여합니다7. DeepSeek의 비용 효율성은 더 넓은 AI 채택에 영향을 미치고 AI 연구실 간의 "가격 경쟁" 가능성을 열어줍니다7.

### 모델 경량화 발전

DeepSeek은 모델 증류를 통해 더 작고 효율적인 버전의 모델(예: 15억, 70억, 700억 파라미터 모델)을 개발하는 데 주력합니다4. 이러한 파생 모델은 파라미터 수가 훨씬 적음에도 불구하고 강력한 성능을 달성하여 소비자용 GPU, iPhone 및 M2 Mac과 같은 리소스 제약이 있는 장치에 배포하는 데 적합합니다4. 모델 경량화는 소규모 팀과 조직이 고급 LLM을 활용할 수 있도록 AI 접근성 및 민주화에 영향을 미칩니다5.

DeepSeek의 효율적인 학습 및 아키텍처에 대한 집중이 LLM 기능의 핵심 영역에서 상당한 결과를 가져왔다는 점을 시사합니다. 벤치마크 점수는 DeepSeek 모델이 저렴할 뿐만 아니라 성능도 뛰어나다는 것을 나타냅니다. 이는 최고 수준의 성능을 얻으려면 막대한 컴퓨팅 자원과 독점 모델이 필요하다는 개념에 도전합니다. 낮은 추론 비용과 경량 모델의 가용성을 결합하여 DeepSeek은 엣지 장치 및 리소스가 제한된 설정을 포함한 광범위한 애플리케이션 및 환경에서 고급 AI 배포를 가능하게 하는 주요 업체로 자리매김하고 있습니다. AI를 더 저렴하게 만들고 다양한 하드웨어 플랫폼에서 접근 가능하게 함으로써 DeepSeek은 AI 기술의 민주화에 기여하고 있으며, 이는 더 많은 일상적인 애플리케이션에 AI가 통합되고 더 넓은 범위의 사용자를 지원할 수 있게 합니다. 오픈 소스 접근 방식과 효율성에 대한 집중은 주요 AI 업체들에게 비용을 절감하고 잠재적으로 모델의 투명성과 접근성을 높이도록 압력을 가하여 전반적으로 더 경쟁적이고 혁신적인 LLM 환경을 조성할 수 있습니다.

## 결론: LLM 미래를 형성하는 DeepSeek의 역할

DeepSeek은 훈련 및 추론 비용의 상당한 절감, 추론 및 코딩 작업에서 비교 가능하거나 우수한 성능, 오픈 소스 원칙 및 모델 경량화에 대한 강력한 약속을 통해 LLM 분야에서 중요한 성과를 거두었습니다. DeepSeek의 혁신은 고급 AI에 대한 접근성을 민주화하고 더욱 경쟁적이고 혁신적인 LLM 생태계를 조성하는 데 잠재적인 영향을 미칩니다. 효율성, 접근성 및 오픈 소스와 독점적 접근 방식 간의 균형 영역에서 DeepSeek은 LLM 개발 방향에 지속적으로 영향을 미칠 가능성이 높으며, LLM의 지속적인 발전에 중요한 기여자이자 혁신적인 파괴자로서의 역할을 확고히 하고 있습니다.

#### Works cited

1. DeepSeek vs. OpenAI: Comparing the New AI Titans | DataCamp, accessed March 18, 2025, <https://www.datacamp.com/blog/deepseek-vs-openai>

2. OpenAI vs DeepSeek Cost Comparison - BytePlus, accessed March 18, 2025, <https://www.byteplus.com/en/topic/385484>

3. DeepSeek R1 vs OpenAI O3 vs Gemini 2 Pro — The Best AI for ..., accessed March 18, 2025, [https://medium.com/@devashish\_m/deepseek-r1-vs-openai-o3-vs-gemini-2-pro-the-best-ai-for-customer-service-executives-46f6c5d8c83a](https://medium.com/%40devashish_m/deepseek-r1-vs-openai-o3-vs-gemini-2-pro-the-best-ai-for-customer-service-executives-46f6c5d8c83a)

4. Open-Source DeepSeek R1 LLM Matches OpenAI's o1 for a Fraction of the Cost, accessed March 18, 2025, <https://www.techeblog.com/china-open-source-deepseek-r1-openai-o1/>

5. What is open-source AI and how could DeepSeek change the industry?, accessed March 18, 2025, <https://www.weforum.org/stories/2025/02/open-source-ai-innovation-deepseek/>

6. Taking Stock of the DeepSeek Shock | FSI - Cyber Policy Center - Stanford University, accessed March 18, 2025, <https://cyber.fsi.stanford.edu/publication/taking-stock-deepseek-shock>

7. DeepSeek: A Game Changer in AI Efficiency? | Bain & Company, accessed March 18, 2025, <https://www.bain.com/insights/deepseek-a-game-changer-in-ai-efficiency/>

8. DeepSeek R1 vs OpenAI o1: Installation, Features, Pricing - Cody, accessed March 18, 2025, <https://meetcody.ai/blog/deepseek-r1-open-source-installation-features-pricing/>

9. www.datacamp.com, accessed March 18, 2025, <https://www.datacamp.com/blog/deepseek-vs-openai#:~:text=What's%20the%20main%20difference%20between,%2460.00%20per%201M%20tokens%20respectively.>

10. OpenAI O3-Mini vs. Deepseek-R1 – Cost, Code Quality & Reasoning Tested - Reddit, accessed March 18, 2025, <https://www.reddit.com/r/ChatGPT/comments/1ikel76/openai_o3mini_vs_deepseekr1_cost_code_quality/>

11. DeepSeek unveils cost-efficient AI model to rival OpenAI - Tech in Asia, accessed March 18, 2025, <https://www.techinasia.com/news/deepseek-unveils-cost-efficient-ai-model-to-rival-openai>

12. DeepSeek-V3 Redefines LLM Performance and Cost Efficiency - DeepLearning.AI, accessed March 18, 2025, <https://www.deeplearning.ai/the-batch/deepseek-v3-redefines-llm-performance-and-cost-efficiency/>

13. What is DeepSeek AI? Is it safe? Here's everything you need to know | ZDNET, accessed March 18, 2025, <https://www.zdnet.com/article/what-is-deepseek-ai-is-it-safe-heres-everything-you-need-to-know/>

14. DeepSeek's LLM: Disrupting the AI Landscape with Low Inference Costs - Marvin Labs, accessed March 18, 2025, <https://www.marvin-labs.com/blog/deepseek-bringing-down-the-ai-cost-curve/>

15. DeepSeek Vs OpenAI: A comparative analysis of LLM development and cost efficiency, accessed March 18, 2025, [https://medium.com/@nrgore1/deepseek-vs-openai-a-comparative-analysis-of-llm-development-and-cost-efficiency-a8534f32c9a8](https://medium.com/%40nrgore1/deepseek-vs-openai-a-comparative-analysis-of-llm-development-and-cost-efficiency-a8534f32c9a8)

16. DeepSeek R-1 Model Overview and How it Ranks Against OpenAI's o1 - PromptHub, accessed March 18, 2025, <https://www.prompthub.us/blog/deepseek-r-1-model-overview-and-how-it-ranks-against-openais-o1>

17. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, accessed March 18, 2025, <https://arxiv.org/html/2501.12948v1>

18. DeepSeek R1 and R1-Zero Explained - The Hundred-Page Language Models Book, accessed March 18, 2025, <https://thelmbook.com/articles/#!./DeepSeek-R1.md>

19. DeepSeek R1: All you need to know - Fireworks AI, accessed March 18, 2025, <https://fireworks.ai/blog/deepseek-r1-deepdive>

20. DeepSeek and the Future of LLMs: Why MongoDB's LLM-agnostic Approach Matters, accessed March 18, 2025, <https://www.mongodb.com/developer/products/atlas/deepseek_rag_mongodb/>

21. GRPO and DeepSeek-R1-Zero. Table of Contents | by Shakti Wadekar | Feb, 2025, accessed March 18, 2025, <https://shaktiwadekar.medium.com/grpo-and-deepseek-r1-zero-9e81f15c6ba2>

22. The DeepSeek Series: A Technical Overview - Martin Fowler, accessed March 18, 2025, <https://martinfowler.com/articles/deepseek-papers.html>

23. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning - The Wire China, accessed March 18, 2025, <https://www.thewirechina.com/wp-content/uploads/2025/01/DeepSeek-R1-Document.pdf>

24. DeepSeek-R1 vs ChatGPT-4o: Analyzing Performance Across Key Metrics. | by Bernard Loki "AI VISIONARY" | Feb, 2025 | Medium, accessed March 18, 2025, [https://medium.com/@bernardloki/deepseek-r1-vs-chatgpt-4o-analyzing-performance-across-key-metrics-2225d078c16c](https://medium.com/%40bernardloki/deepseek-r1-vs-chatgpt-4o-analyzing-performance-across-key-metrics-2225d078c16c)

25. deepseek-ai/DeepSeek-R1 - GitHub, accessed March 18, 2025, <https://github.com/deepseek-ai/DeepSeek-R1>

26. DeepSeek R1: Pioneering Open-Source 'Thinking Model' and Its Impact on the LLM Landscape - UNU Campus Computing Centre, accessed March 18, 2025, <https://c3.unu.edu/blog/deepseek-r1-pioneering-open-source-thinking-model-and-its-impact-on-the-llm-landscape>

27. DeepSeek's impact on thoracic surgeons' work patterns—past, present and future - PMC, accessed March 18, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC11898358/>

28. DeepSeek R1 Explained: Chain of Thought, Reinforcement Learning, and Model Distillation | by Tahir | Jan, 2025 | Medium, accessed March 18, 2025, [https://medium.com/@tahirbalarabe2/deepseek-r1-explained-chain-of-thought-reinforcement-learning-and-model-distillation-0eb165d928c9](https://medium.com/%40tahirbalarabe2/deepseek-r1-explained-chain-of-thought-reinforcement-learning-and-model-distillation-0eb165d928c9)

29. DeepSeek: Everything you need to know about this new LLM in one place - Daily.dev, accessed March 18, 2025, <https://daily.dev/blog/deepseek-everything-you-need-to-know-about-this-new-llm-in-one-place>

30. DeepSeek and the Power of Mixture of Experts (MoE) - DEV Community, accessed March 18, 2025, <https://dev.to/sayed_ali_alkamel/deepseek-and-the-power-of-mixture-of-experts-moe-ham>

31. DeepSeek Technical Analysis — (1) Mixture-of-Experts | by Jinpeng Zhang | Jan, 2025, accessed March 18, 2025, <https://dataturbo.medium.com/key-techniques-behind-deepseek-models-10x-efficiency-1-moe-9bd2534987c8>

32. Exploring DeepSeek-R1's Mixture-of-Experts Model Architecture - AI Resources - Modular, accessed March 18, 2025, <https://www.modular.com/ai-resources/exploring-deepseek-r1-s-mixture-of-experts-model-architecture>

33. Everything You Need to Know About DeepSeek's Reasoning-Focused LLM - Ascentt, accessed March 18, 2025, <https://www.ascentt.com/everything-you-need-to-know-about-deepseeks-reasoning-focused-llm/>

34. (PDF) DeepSeek-V3: A High-Performance Mixture-of-Experts Language Model, accessed March 18, 2025, <https://www.researchgate.net/publication/388484212_DeepSeek-V3_A_High-Performance_Mixture-of-Experts_Language_Model>

35. DeepSeek-V3 Technical Report - arXiv, accessed March 18, 2025, <https://arxiv.org/html/2412.19437v1?ref=platformer.news>

36. DeepSeek v3 Review: Performance in Benchmarks & Evals - TextCortex, accessed March 18, 2025, <https://textcortex.com/post/deepseek-v3-review>

37. DeepSeek V3 vs. GPT-4o: Architecture, Performance Comparison | by Novita AI - Medium, accessed March 18, 2025, [https://medium.com/@marketing\_novita.ai/deepseek-v3-vs-gpt-4o-architecture-performance-comparison-5619138be50a](https://medium.com/%40marketing_novita.ai/deepseek-v3-vs-gpt-4o-architecture-performance-comparison-5619138be50a)

38. deepseek-ai/DeepSeek-MoE: DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models - GitHub, accessed March 18, 2025, <https://github.com/deepseek-ai/DeepSeek-MoE>

39. DeepSeek-AI - Papers With Code, accessed March 18, 2025, <https://paperswithcode.com/author/deepseek-ai>

40. DeepSeek V3: The Open-Source AI Revolution - Dirox, accessed March 18, 2025, <https://dirox.com/post/deepseek-v3-the-open-source-ai-revolution>

41. DeepSeek: Bridging Performance and Efficiency in Modern AI | by Nandini Lokesh Reddy, accessed March 18, 2025, [https://medium.com/@nandinilreddy/deepseek-bridging-performance-and-efficiency-in-modern-ai-106181a85693](https://medium.com/%40nandinilreddy/deepseek-bridging-performance-and-efficiency-in-modern-ai-106181a85693)

42. DeepSeek-V3 Technical Report - arXiv, accessed March 18, 2025, <https://arxiv.org/html/2412.19437v1>

43. DeepSeek-V3 Technical Report - arXiv, accessed March 18, 2025, <https://arxiv.org/pdf/2412.19437>

44. How DeepSeek Trained AI 30 Times Cheaper? - Analytics Vidhya, accessed March 18, 2025, <https://www.analyticsvidhya.com/blog/2025/01/how-deepseek-trained-ai-30-times-cheaper/>

45. DeepSeek R1: How Hardware Constraints Led to an AI Innovation Revolution - Meibel, accessed March 18, 2025, <https://www.meibel.ai/post/deepseek-r1-how-hardware-constraints-led-to-an-ai-innovation-revolution>

46. Deepseek-R1: The best Open-Source Model, But how to use it? | by Pankaj Tiwari - Medium, accessed March 18, 2025, <https://medium.com/accredian/deepseek-r1-the-best-open-source-model-but-how-to-use-it-fb0dd28c1557>

47. DeepSeek-R1 now available as a fully managed serverless model in Amazon Bedrock, accessed March 18, 2025, <https://aws.amazon.com/blogs/aws/deepseek-r1-now-available-as-a-fully-managed-serverless-model-in-amazon-bedrock/>

48. DeepSeek wraps up week of AI model disclosures - Tech in Asia, accessed March 18, 2025, <https://www.techinasia.com/news/deepseek-wraps-up-week-of-ai-model-disclosures/amp/>

49. DeepSeek R1 Review: Performance in Benchmarks & Evals - TextCortex, accessed March 18, 2025, <https://textcortex.com/post/deepseek-r1-review>

50. DeepSeek API Docs: Your First API Call, accessed March 18, 2025, <https://api-docs.deepseek.com/>

51. DeepSeek API: A Guide With Examples and Cost Calculations - DataCamp, accessed March 18, 2025, <https://www.datacamp.com/tutorial/deepseek-api>

52. How to Use DeepSeek API for Free: A Step-by-Step Guide - Apidog, accessed March 18, 2025, <https://apidog.com/blog/how-to-use-deepseek-api-for-free/>

53. Mastering the Deepseek API (R1 & V3): A Developer's Guide - Sebastian Petrus - Medium, accessed March 18, 2025, <https://sebastian-petrus.medium.com/deepseek-api-r1-v3-guide-7be11be4be9a>

54. DeepSeek-R1: Features, o1 Comparison, Distilled Models & More | DataCamp, accessed March 18, 2025, <https://www.datacamp.com/blog/deepseek-r1>

55. OpenAi vs DeepSeek Price Comparison | by John Gragson | Jan, 2025 - Medium, accessed March 18, 2025, [https://medium.com/@gragson.john/openai-vs-deepseek-price-comparison-aa199f049bb8](https://medium.com/%40gragson.john/openai-vs-deepseek-price-comparison-aa199f049bb8)

56. How DeepSeek R1 Was Trained: A Cost-Efficient Revolution in AI - BytePlus, accessed March 18, 2025, <https://www.byteplus.com/en/topic/385503>

57. Why DeepSeek V3 is the LLM Everyone's Talking About - Spheron's Blog, accessed March 18, 2025, <https://blog.spheron.network/why-deepseek-v3-is-the-llm-everyones-talking-about>

58. Breaking the AI Mold: China's DeepSeek-R1 pushes local and open AI forward - WorkOS, accessed March 18, 2025, <https://workos.com/blog/deepseek-r1-pushes-local-and-open-ai-forward>