DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

최수빈

연구 배경

LLM(대형 언어 모델)의 한계와 해결책

- 최근 LLM(예: GPT, Claude, Llama)은 빠르게 발전하고 있지만, 추론 능력에서 여전히 한계를 보임.
- OpenAI의 o1 모델은 Chain-of-Thought(CoT) 기반의 추론 확장을 통해 성능을 향상시켰지만, 폐쇄형 모델임.
- DeepSeek-AI는 강화 학습(RL) 을 활용하여 오픈소스 기반으로 LLM의 추론 능력을 개선하는 방법을 연구.

모델	주요 특징
기존 LLM (GPT-4, Claude)	지도 미세 조정(SFT) + RLHF 활용, 폐쇄형
OpenAI o1-1217	CoT 기반 추론 확장, 높은 성능 but 비공개
DeepSeek-R1	순수 RL 기반 추론 향상, 오픈소스 공개

연구 목표

DeepSeek-R1 연구의 핵심 목표는 다음과 같음

- 강화 학습(RL)만으로 추론 능력을 향상시킬 수 있는지 탐구
- 지도 미세 조정(SFT) 없이 학습한 DeepSeek-R1-Zero의 성능 평가
- SFT와 RL을 결합한 DeepSeek-R1 개발 및 성능 개선
- 작은 모델에도 추론 능력을 증류하여 효율적인 AI 모델 설계

최종 목표

- OpenAI o1-1217과 유사한 성능을 내면서도 오픈소스 모델을 제공

방법론 개요 : DeepSeek-R1 학습 과정

DeepSeek-R1-Zero

- SFT 없이 **순수 강화 학습(RL)만으로 학습**
- 놀라운 추론 능력을 보였지만, 가독성 문제 발생

DeepSeek-R1

- **콜드 스타트 데이터(SFT) 추가 + RL 적용**
- OpenAI o1-1217과 비슷한 성능 달성

DeepSeek-R1-Distill (증류 모델 : 원본 모델의 지식을 작은 모델에게 전수)

- **DeepSeek-R1을 작은 모델(1.5B~70B)로 증류**
- **작은 모델에서도 강력한 추론 능력 유지**

기술적 접근법:

- GRPO(Group Relative Policy Optimization) 알고리즘 적용
- 콜드 스타트 데이터를 활용한 사전 미세 조정
- RL 과정에서 보상 모델링 및 거부 샘플링 적용

모델 성능 비교

모델	핵심 특징	장점	단점
DeepSeek-R1-Zero	- SFT 없이 순수 RL 적용 - 자기 진화(self-evolution)	• 놀라운 추론 능력 • Pass@1 점수 급상승	• 가독성 문제 언어 혼합 현상
DeepSeek-R1	- 콜드 스타트 데이터 추가 - RL과 SFT 결합 - 추론 중심 강화 학습 적용	• OpenAI o1-1217과 유사한 성능 • 가독성 개선 • 다국어 지원 향상	• 소프트웨어 엔지니어링 성능 개선 필요
DeepSeek-R1-Distill	- 작은 모델(1.5B~70B)로 증류 - RL 없이 SFT만 적용	• 작은 모델에서도 강한 성능 유지 • 효율적인 AI 모델 설계	• RL을 포함하면 더 높은 성능 가능

요약 :
DeepSeek-R1의 발전 과정은 **순수 RL → SFT+RL → 증류** 단계로 진행되며, **추론 능력과 가독성**이 점점 향상됨.

모델 성능 비교 - 2

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	1820
OpenAI-o1-0912	74.4	83.3	94.8	77.3	63.4	1843
DeepSeek-R1-Zero	71.0	86.7	95.9	73.3	50.0	1444

Table 2 | Comparison of DeepSeek-R1-Zero and OpenAI o1 models on reasoning-related benchmarks.

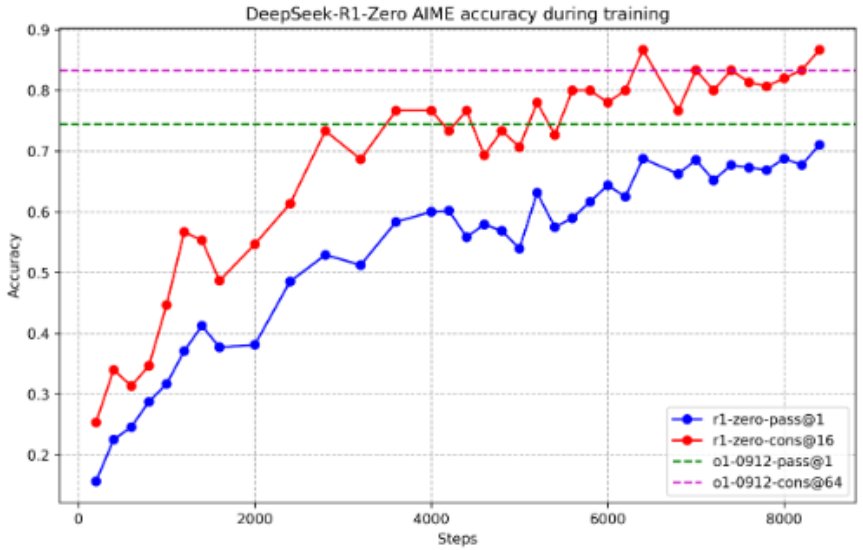

Figure 2 | AIME accuracy of DeepSeek-R1-Zero during training. For each question, we sample 16 responses and calculate the overall average accuracy to ensure a stable evaluation.

DeepSeek-R1-Zero vs. OpenAI o1 모델 비교

- DeepSeek-R1-Zero는 수학적 추론 능력이 매우 뛰어나며, 학습이 진행될수록 성능이 향상됨
- 하지만 Pass@1 성능이 OpenAI o1-0912보다 낮아, 답을 한 번에 맞히는 능력은 개선이 필요함
- LiveCode Bench 성능이 낮아 소프트웨어 엔지니어링 능력이 부족한 점도 보완해야 함

모델 성능 비교 – 3 : 다른 모델들과의 비교

	Benchmark (Metric)	Claude-3.5-Sonnet-1022	GPT-4o 0513	DeepSeek V3	OpenAI o1-mini	OpenAI o1-1217	DeepSeek R1
	Architecture	-	-	MoE	-	-	MoE
	# Activated Params	-	-	37B	-	-	37B
	# Total Params	-	-	671B	-	-	671B
English	MMLU (Pass@1)	88.3	87.2	88.5	85.2	**91.8**	90.8
	MMLU-Redux (EM)	88.9	88.0	89.1	86.7	-	**92.9**
	MMLU-Pro (EM)	78.0	72.6	75.9	80.3	-	**84.0**
	DROP (3-shot F1)	88.3	83.7	91.6	83.9	90.2	**92.2**
	IF-Eval (Prompt Strict)	**86.5**	84.3	86.1	84.8	-	83.3
	GPQA Diamond (Pass@1)	65.0	49.9	59.1	60.0	**75.7**	71.5
	SimpleQA (Correct)	28.4	38.2	24.9	7.0	**47.0**	30.1
	FRAMES (Acc.)	72.5	80.5	73.3	76.9	-	**82.5**
	AlpacaEval2.0 (LC-winrate)	52.0	51.1	70.0	57.8	-	**87.6**
	ArenaHard (GPT-4-1106)	85.2	80.4	85.5	92.0	-	**92.3**
Code	LiveCodeBench (Pass@1-COT)	38.9	32.9	36.2	53.8	63.4	**65.9**
	Codeforces (Percentile)	20.3	23.6	58.7	93.4	**96.6**	96.3
	Codeforces (Rating)	717	759	1134	1820	**2061**	2029
	SWE Verified (Resolved)	**50.8**	38.8	42.0	41.6	48.9	49.2
	Aider-Polyglot (Acc.)	45.3	16.0	49.6	32.9	**61.7**	53.3
Math	AIME 2024 (Pass@1)	16.0	9.3	39.2	63.6	79.2	**79.8**
	MATH-500 (Pass@1)	78.3	74.6	90.2	90.0	96.4	**97.3**
	CNMO 2024 (Pass@1)	13.1	10.8	43.2	67.6	-	**78.8**
Chinese	CLUEWSC (EM)	85.4	87.9	90.9	89.9	-	**92.8**
	C-Eval (EM)	76.7	76.0	86.5	68.9	-	**91.8**
	C-SimpleQA (Correct)	55.4	58.7	**68.0**	40.3	-	63.7

Table 4 | Comparison between DeepSeek-R1 and other representative models.

Deepseek-R1의 특징

강점:

- 수학적 추론 능력 최강 (MATH-500, AIME 2024 최고점)
- 중국어 이해 및 응답 품질 최상위 (CLUEWSC, C-Eval)
- AlpacaEval 2.0(NLP 성능 평가) & ArenaHard(AI 모델 응답 품질 평가)최고점 → 전반적인 AI 응답 품질 우수

약점:

- 코딩 능력 부족 (LiveCodeBench, Codeforces 성능이 OpenAI 모델보다 낮음)
- 영어 MMLU 성능이 OpenAI o1-1217보다 떨어짐
- GPQA Diamond (복잡한 질문 응답 정확도)도 OpenAI보다 낮음

요약 : DeepSeek-R1은 "수학과 중국어에 강한 모델이지만 코딩과 복잡한 영어 문제 해결에는 약점이 있음.

실험 결과와 평가

벤치마크	DeepSeek-R1	OpenAI o1-1217
AIME 2024(수학적 추론)	79.8%	79.6%
MATH-500(수학)	97.3%	97.1%
LiveCodeBench(코딩)	57.2%	56.8%

- 연구 목표대로 DeepSeek-R1은 OpenAI o1-1217과 유사한 성능 달성
- 수학, 논리 추론에서 강력한 결과

논의 및 한계

강점:

- 강화 학습이 **추론 능력 향상에 효과적**
- OpenAI o1-1217과 **비슷한 성능 달성**

한계점:

- RL만 적용한 모델(DeepSeek-R1-Zero)의 **가독성 문제**
- **프롬프트 민감도 높음** (Few-shot에서 성능 저하)
- 소프트웨어 엔지니어링 작업에서 성능 개선 필요

향후 연구 방향:

- 소프트웨어 엔지니어링 작업 성능 개선
- 언어 혼합 문제 해결
- 긴 문맥 이해력 향상

결론

- DeepSeek-R1-Zero는 SFT 없이도 RL만으로 강력한 추론 능력을 발휘
- DeepSeek-R1은 SFT + RL을 활용하여 성능을 극대화
- 증류 기법을 통해 작은 모델에서도 강력한 성능 유지
- 연구 커뮤니티를 위해 1.5B~70B 크기의 모델을 오픈소스로 공개

- 결론: DeepSeek-R1은 오픈소스 기반 LLM의 새로운 가능성을 열었으며, 강화 학습이 추론 능력을 향상시킬 수 있음을 증명