대규모 언어모델의 자기성찰 능력 향상을 위한 피드백 최적화 기법

김준석, 정교민
서울대학교
kim.junseok@snu.ac.kr, kjung@snu.ac.kr

Optimizing Feedback to Improve Self-Reflection in Large Language Models

Kim Jun Seok, Jung Kyo Min
Interdisciplinary Program in Artificial Intelligence, Seoul National Univ.

요 약

본 논문은 대규모 언어모델의 낮은 자기성찰(self-reflection) 성능의 원인을 피드백 품질 부족에서 찾고, 이를 개선하여 성능을 향상시키는 방법을 제안한다. LLM을 활용해 초기 답변, 피드백, 그리고 피드백 이후의 답변을 생성하고, 답변의 정답률을 기반으로 피드백 품질을 정량화한 뒤, 이를 바탕으로 선호도 데이터를 구축하여 DPO 방식으로 LLM을 학습시켰다. 학습된 LLM은 자기성찰 후 성능 향상에 기여했으며, 피드백 품질이 성찰 기반 문제 해결 성능 향상에 중요한 역할을 기여함을 확인하였다. 또한 트리 구조를 활용한 데이터 생성 방식으로 특정 도메인에 한정되지 않고, 다양한 과제에 적용 가능한 일반화된 자기성찰 향상 기법을 제시한다.

Ⅰ. 서 론

본 논문에서는 자기성찰(self-reflection) 능력을 향상시키기 위한 새로운 접근 방식을 제안한다. 자기성찰이란 문제를 푸는 과정에서 자신의 답변이 정답인지 여부를 스스로 평가하고, 그 결과에 따라 답변을 수정하는 과정을 의미한다. [1] 이는 인간에게 있어 자연스러운 사고 과정이지만, 기존의 대규모 언어모델(Large Language Models, LLMs)은 이러한 능력을 본질적으로 갖추고 있지 않다. 실제로 LLM이 자기성찰을 수행할 경우, 오히려 성능이 저하된다는 사실이 여러 선행 연구를 통해 밝혀져 왔다. [2]

이러한 문제를 해결하기 위한 시도로, 최근에는 LLM 간의 찬반 토론을 통해 자기성찰을 유도하는 방식이 제안되었다. [3] 그러나 이 방식은 높은 연산 비용을 요구하며, 반박을 생성하는 모델의 품질이 낮을 경우 오히려 잘못된 방향으로 자기성찰이 이루어질 수 있다는 한계가 있다.

이에 본 연구에서는 자기성찰 과정에서 생성되는 피드백의 품질을 높이는 데 초점을 맞추었다. 구체적으로는 LLM을 이용하여 초기 답변, 해당 답변에 대한 피드백, 그리고 피드백을 반영한 후속 답변을 순차적으로 생성함으로써 피드백 품질과 관련된 데이터셋을 구축하였다. 이후, 이 데이터셋을 기반으로 강화학습을 적용하여 초기 답변에 대해 보다 정확하고 유익한 피드백을 생성할 수 있도록 LLM을 학습시켰다.

양질의 피드백은 자기성찰의 효과를 극대화하는 핵심 요소이며, 본 연구는 이러한 피드백의 질적 향상을 통해 LLM의 자기성찰 기반 문제 해결 능력을 보다 효율적으로 강화하는 것을 목표로 한다.

Ⅱ. 본론

본 논문에서 제안하는 방법론은 크게 두 단계로 나뉜다. 첫 번째 단계에서는 피드백의 품질을 정량적으로 측정할 수 있는 데이터셋을 생성하고, 두 번째 단계에서는 이 데이터를 활용하여 LLM이 더 나은 피드백을 생성할 수 있도록 학습시킨다.

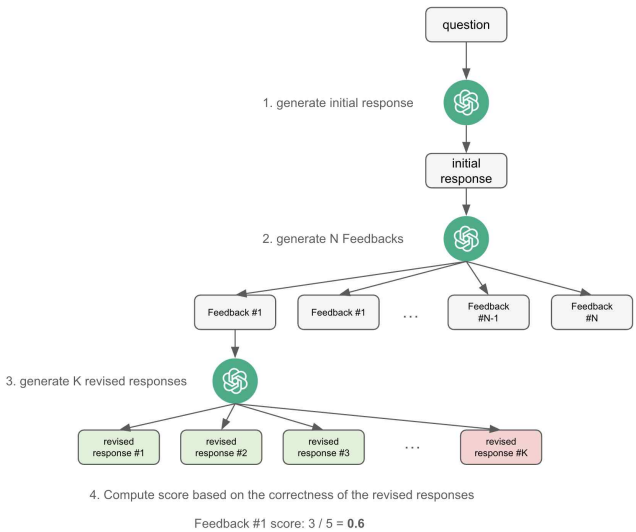

그림 1 피드백 품질 관련된 데이터 생성 방식

그림 1은 피드백 품질에 관련된 데이터를 생성하는 방식에 대해서 요약하고 있다. 먼저, 주어진 문제에 대해 LLM의 temperature 값을 1로 설정하여 초기 답변에 대한 다양한 피드백과 추후 답변을 얻어낼 수 있다. temperature 값을 높게 설정함으로써 보다 다

양한 방식의 답변을 유도할 수 있으며, 이는 후속 피드백과 자기성찰 답변의 다양성을 확보하는 데 도움이 된다. 생성된 초기 답변에 대해서, 동일한 LLM을 사용하여 n개의 피드백을 생성한다. 이 피드백은 각 초기 답변의 오류를 지적하거나 개선 방향을 제시하는 텍스트로 구성된다.

그 다음 단계에서는, 생성된 각 피드백에 대해 다시 LLM을 활용하여 n개의 자기성찰 이후 k개의 답변을 생성한다. 즉, 하나의 피드백이 있을 때, 이를 반영한 후속 답변을 여러 개 생성하여 피드백의 효과를 측정하는 방식이다. 이렇게 생성된 자기성찰 답변 중에서 정답을 맞춘 답변의 개수를 세고, 이를 k로 나눈 값을 해당 피드백의 품질 점수로 정의한다. 예를 들어, 하나의 피드백으로부터 생성된 5개의 자기성찰 답변 중 3개가 정답일 경우, 이 피드백의 점수는 0.6이 된다.

이러한 방식으로, 문제와 초기 답변, 피드백, 그리고 해당 피드백의 정답 유도율에 기반한 점수로 구성된 데이터셋을 구축할 수 있다. 이후 이 데이터셋에서 피드백 간의 점수를 비교함으로써 어떤 피드백이 더 나은지를 판단할 수 있으며, 이 비교 결과를 활용하여 선호도에 따른 데이터 쌍을 구성한다. 예를 들어, 두 개의 피드백 중 하나의 점수가 더 높다면, 그 피드백을 우선적으로 선호하는 데이터 쌍을 만들 수 있다.

이렇게 생성된 데이터를 바탕으로, LLM을 Direct Preference Optimization(DPO) 방식으로 학습시킨다. [4] DPO는 모델이 더 나은 출력을 생성하도록 학습하는 방식으로, 본 연구에서는 초기 답변과 피드백을 입력으로 받아 더 우수한 피드백을 출력하는 LLM을 만드는 데 사용된다. 학습된 모델은 동일한 초기 답변에 대해 더 높은 품질의 피드백을 생성할 수 있으며, 이는 자기성찰을 거친 최종 답변의 정확도를 향상시키는 데 기여한다.

모델	데이터셋	자가성찰 방법론	
		일반 모델	학습된 모델
llama-3.2-3B-Instruct	GSM8K	53.29	71.57 **(+18.28)**
	CSQA	36.12	64.54 **(+28.42)**
	CoinFlip	27.35	38 **(+10.65)**

표 1 피드백 생성 시 학습된 모델의 성능 비교

제안한 방법론의 유효성을 검증하기 위해, 기존 연구에서 널리 사용된 세 가지 추론 벤치마크에 대해 실험을 수행하였다. 특히 벤치마크 선정 기준으로는 수학, 상식, 기호적 추론 능력을 다루는 다양한 도메인에 대해서 평가하고자 하였다. 실험 모델로는 보편적으로 일반 컴퓨터에서도 사용할 수 있는 3B 크기의 LLM을 사용하였다. [5] 학습된 LLM을 사용하여 피드백을 생성하고, 이를 자기성찰에 반영한 후 최종 답변의 성능을 평가하였다. 실험 결과는 표 1에 제시되어 있으며, 학습 이전의 기존 LLM과 비교했을 때 학습된 LLM을 사용할 경우 피드백 이후의 성능이 일관되게 향상됨을 확인할 수 있었다. 이는 피드백의 품질이 자기성찰 성능에 직결되며, 본 연구에서 제안한 피드백 품질 향상 기법이 실제로 문제 해결 능력 향상에 효과적이라는 것을 보여준다.

Ⅲ. 결론

본 연구에서는 기존 LLM의 자기성찰 성능이 낮은 원인 중 하나로 피드백의 품질 부족에 주목하고, 이를 개선함으로써 자기성찰 이후의 최종 성능을 유의미하게 향상시킬 수 있음을 보였다. 단순히 자기성찰을 수행하는 것만으로는 성능 저하가 발생할 수 있으나, 본 논문에서 제안한 방식처럼 피드백의 품질을 정량적으로 평가하고 이를 학습에 반영할 경우, 성찰 과정 자체가 실질적인 도움이 될 수 있다는 점을 실험을 통해 입증하였다. 또한, 기존의 자기성찰 능력 향상 기법들이 주로 수학적 추론과 같은 특정 도메인에 집중되어 있었던 반면, 본 연구는 트리 구조를 활용한 데이터 생성 방식을 도입함으로써 도메인에 종속되지 않는 자기성찰 향상 방법론을 제안하였다. 이를 통해 수학 문제뿐만 아니라 보다 일반적인 자연어 처리 과제에서도 자기성찰 기반 성능 향상이 가능함을 확인하였다. 향후에는 본 연구에서 제안한 방법론을 다양한 태스크와 모델 아키텍처에 확장하여, LLM의 보편적인 자기성찰 능력 향상에 기여할 수 있을 것으로 기대한다.

ACKNOWLEDGMENT

이 논문은 *2025*년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 **[NO.RS-2021-II211343,인공지능대학원지원(서울대학교)]**

참 고 문 헌

[1] Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., ... & Clark, P. (2023). Self-refine: Iterative refinement with self-feedback.*Advances in Neural Information Processing Systems,36*, 46534-46594.

[2] Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., & Zhou, D. (2023). Large language models cannot self-correct reasoning yet.*arXiv preprint arXiv:2310.01798*.

[3] Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., ... & Tu, Z. (2024, November). Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. In*Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*(pp. 17889-17904).

[4] Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model.*Advances in Neural Information Processing Systems,36*, 53728-53741.

[5] Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., ... & Vasic, P. (2024). The llama 3 herd of models.*arXiv preprint arXiv:2407.21783*.