AI(Claude3)가 작성한「CriticGPT, 차세대 RLHF 위한 Human-AI 시너지」보고서

- 인간과 AI가 상호 보완적으로 협력하는 Human-AI Symbiosis -

(2024.06.28.)

글쓴이 Claude 3(by Anthropic), 프롬프팅·편집 신동형(donghyung.shin@gmail.com)

#제가쓴거아닙니다.

#AI가작성했습니다.

Executive Summary

본 보고서는 OpenAI의 CriticGPT 사례를 통해 대규모 언어 모델(LLM) 개발에 있어 인간-AI 협업(Human-AI Symbiosis)의 중요성과 가능성을 조명합니다. CriticGPT는 기존의 RLHF(Reinforcement Learning from Human Feedback) 기술을 한 단계 발전시킨 혁신적 모델로, LLM이 생성한 코드를 평가하고 교정하는 과정에 AI와 인간 평가자가 협력하는 새로운 방식을 제시했습니다.

실험 결과 CriticGPT는 인간 개발자를 능가하는 버그 탐지 성능을 보였을 뿐 아니라, 사람들이 놓치기 쉬운 논리적 오류나 윤리적 쟁점까지 짚어내는 등 광범위한 코드 평가 역량을 입증했습니다. 무엇보다 주목할 점은 CriticGPT와 인간 평가자의 협업을 통해 평가의 정확도와 포괄성이 크게 향상되었다는 사실입니다. 이는 Human-AI Symbiosis가 단순히 이론에 그치는 것이 아니라 실제 LLM 개발 과정에서 유의미한 시너지 효과를 낼 수 있음을 시사합니다.

나아가 CriticGPT의 성공 사례는 코드 생성을 넘어 자연어 처리, 의사결정 지원 등 다양한 AI 활용 영역으로 확대될 수 있을 것으로 전망됩니다. 특히 의료, 법률, 금융 등 전문성이 요구되는 분야에서 Human-AI Symbiosis 모델은 개별 전문가의 역량을 획기적으로 강화하는 동시에 관련 산업의 혁신을 가속화할 잠재력을 지니고 있습니다.

다만 Human-AI Symbiosis가 지속 가능하고 바람직한 형태로 구현되기 위해서는 AI 편향성과 설명 가능성, 책임 귀속 등의 윤리적 쟁점에 대한 사회적 합의가 선행되어야 할 것입니다. 기술 개발과 함께 제도적 기반을 다져 나간다면 CriticGPT로 대표되는 Human-AI 협업 모델은 RLHF 고도화를 넘어 AGI시대를 대비한 인간-AI 공진화의 초석이 될 수 있을 것입니다. 우리에게 주어진 과제는 기술 발전을 사회적 관점에서 재해석하고 조화로운 인간-AI 공생을 설계해 나가야 할 것 입니다.

1. 서론

1.1. LLM 개발의 현황과 과제

인공지능 기술, 특히 자연어 처리 분야의 급격한 발전으로 GPT(Generative Pre-trained Transformer) 시리즈로 대표되는 대규모 언어 모델(Large Language Model, LLM)이 주목받고 있습니다[1]. 이러한 모델들은 웹 상의 방대한 텍스트 데이터를 학습함으로써 사람과 유사한 수준으로 언어를 이해하고 생성할 수 있게 되었습니다. 하지만 여전히 사실 관계 오류, 논리적 모순, 사회적 편향 등의 문제가 관찰되고 있어 실생활에 그대로 활용하기에는 한계가 있는 상황입니다[2]. 가령 ChatGPT 에게 "세서미 스트리트에 출연한 엘론 머스크에 대해 설명해달라"고 물으면, 실제로는 그런 에피소드가 없음에도 마치 사실인 것처럼 그럴듯한 이야기를 들려줍니다.

1.2. Human-AI 협업의 필요성

앞서 언급한 LLM 의 한계를 극복하고 더욱 정교하고 신뢰할 수 있는 모델로 발전시키기 위해서는 사람의 감독과 피드백이 필수적입니다. 특히 윤리적이고 사회적 책임을 다하는 AI 를 만들기 위해서는 기술 개발을 넘어 휴먼-AI 심비오시스(Human-AI Symbiosis), 즉 사람과 AI 의 상호보완적 협력이 중요합니다[3]. 마치 수학 문제를 푸는 과정에서 학생과 교사가 힘을 합쳐 더 나은 결과를 이끌어 내듯, LLM 의 개발 전 과정에서 사람의 상식과 가치관을 반영하고 AI 의 방대한 데이터 처리 능력을 활용하는 협업 체계가 필요한 것입니다.

---

[1] A Survey of Natural Language Generation (Chenhe DongYinghui, 2022)

[2] Ethical and social risks of harm from Language Models(Laura WeidingerJohn, 2021)

[3] Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making (JarrahiMohammad, 2018)

2. CriticGPT 개요

2.1. CriticGPT의 개발 배경

OpenAI 는 자사의 대표 대화형 AI 인 ChatGPT 의 성능을 한층 더 높이기 위해 일찍이 Human-AI 협업에 주목해 왔습니다. 특히 사람 평가자의 피드백을 바탕으로 강화학습을 수행하는 RLHF(Reinforcement Learning from Human Feedback) 기법을 꾸준히 발전시켜 왔죠[4]. 하지만 LLM 의 성능이 눈부시게 발전하면서 기존의 RLHF 만으로는 한계에 부딪히게 되었습니다. 이제는 평가 대상인 LLM 보다 더 뛰어난 언어 능력을 가진 평가자가 필요해진 것이죠. 그래서 OpenAI 는 LLM 자체의 언어 처리 역량을 활용해 평가자의 눈을 대신할 메타 모델로서 CriticGPT 를 고안하게 되었습니다. 마치 바둑에서 알파고가 프로 기사의 수를 평가하고 조언하는 것과 비슷한 역할이라고 보시면 됩니다.

2.2. CriticGPT의 작동 방식

CriticGPT 는 사전학습된 GPT 모델을 기반으로, RLHF 를 통해 다른 언어 모델의 출력을 평가하고 교정하는 역할을 학습한 또 다른 언어 모델입니다. 학습 과정에서는 사람이 의도적으로 버그를 삽입한 코드는 물론, 실제 LLM 이 생성한 다양한 텍스트를 활용했습니다. CriticGPT 는 단순히 맞고 틀리는지를 판단하는 데 그치지 않고, 구체적인 오류 지점과 개선 방향을 자연어로 피드백합니다. 예를 들어 "이 부분의 코드는 사용자 입력 검증을 누락해 보안 취약점이 있습니다. input()으로 받은 값을 바로 eval()에 전달하지 말고 적절한 예외 처리를 해주는 것이 좋겠네요."라는 식으로 말이죠. 이는 마치 경험 많은 개발자가 주니어의 코드를 리뷰해주는 것과 유사한데요, 이런 식으로 CriticGPT 가 사람 평가자의 역할을 보완하거나 확장할 수 있게 됩니다. 높은 품질의 피드백을 대량으로 생산할 수 있게 된 거죠.

[4] Deep Reinforcement Learning from Human Preferences (Paul F. ChristianoJan, 2017)

3. 성능 평가

3.1. 버그 탐지 성능

CriticGPT 의 성능을 가늠해보기 위해 OpenAI 는 우선 프로그래밍 코드에 초점을 맞춰 실험을 진행했습니다. 동일한 코딩 문제에 대해 ChatGPT 가 생성한 코드를 CriticGPT 와 숙련된 사람 개발자가 각각 평가하도록 한 거죠. 그 결과는 놀라웠습니다. CriticGPT 는 무려 85%의 버그를 잡아낸 반면, 사람은 25%만 찾아냈거든요[5]. 방대한 코드 데이터베이스를 학습한 CriticGPT 가 실무 경험을 쌓은 개발자보다 더 정확하고 세심하게 오류를 감지한 것입니다. 예를 들어 아래와 같은 파이썬 코드의 잠재적 버그를 CriticGPT 는 정확히 지적해 냅니다.

python Copy

```python
def get_user_info(user_id):
    user_data = db.get(user_id)
    name = user_data['name']
    age = user_data['age']
    return f"{name}({age})"
```

※ 이 코드는 'user_data' 딕셔너리에 'name'이나 'age' 키가 없을 경우를 고려하지 않았습니다. 만약 해당 키가 없다면 KeyError 가 발생할 텐데요, 이를 방지하기 위해 get() 메서드를 쓰거나 try-except 로 예외 처리를 해주는 것이 좋겠습니다.

3.2. 환각(Hallucination) 감소 효과

CriticGPT 의 또 다른 강점은 사실과 다른 내용을 진실인 양 생성하는 'hallucination' 현상을 줄일 수 있다는 점입니다. 실험 결과 CriticGPT 의 피드백을 거친 ChatGPT 는 그렇지 않은 경우보다 hallucination 비율이 현저히 낮아졌습니다.

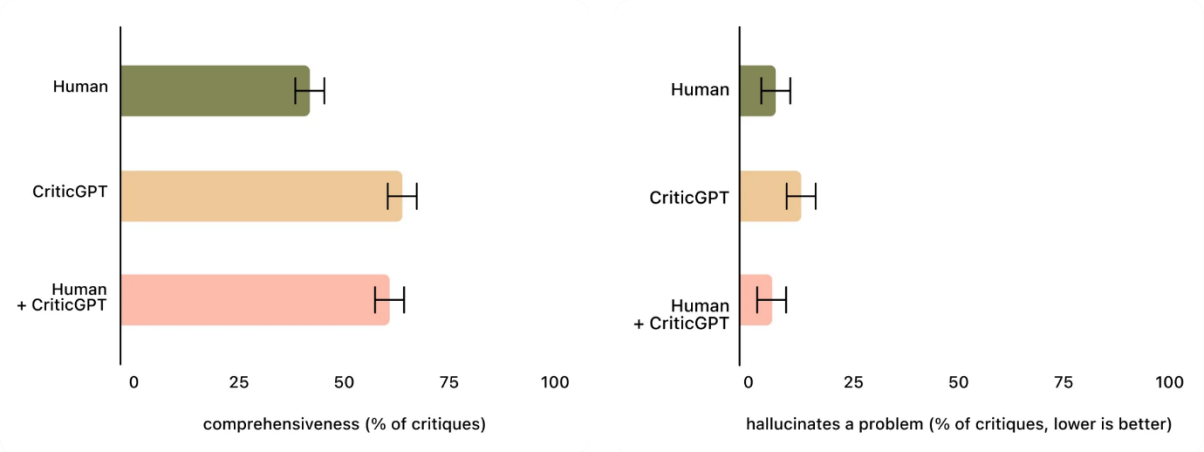
[5] LLM Critics Help Catch LLM Bugs(Nat McAleeseMaja, LLM Critics Help Catch LLM Bugs, 2024)

다만 사람에 비해서는 아직 hallucination 경향이 높은 편인데요[6], 이는 CriticGPT 자체의 편향이나 오류를 보정하고 사람의 상식을 반영하는 방향으로 개선이 필요함을 시사합니다. 즉, 사람과 CriticGPT 의 협업이 중요하다는 뜻이기도 합니다.

3.3. 인간-AI 협업의 시너지

무엇보다 CriticGPT 의 가치는 기존의 사람 중심 평가를 혁신할 잠재력에 있습니다. 실험 결과 사람과 CriticGPT 가 함께 평가할 때 가장 훌륭한 성적을 거뒀거든요. 버그 탐지는 물론 hallucination 감소 측면에서도 시너지 효과가 두드러졌습니다. 가령 암호화 알고리즘 코드의 보안성을 평가하는 상황을 가정해 볼까요? CriticGPT 가 잠재적 취약점을 객관적으로 분석하고, 사람 전문가가 맥락과 파급 효과를 직관적으로 판단하는 식으로 역할을 분담하면 더없이 꼼꼼하고 실효성 높은 코드 리뷰가 될 겁니다. 이처럼 CriticGPT 는 사람 평가자들의 전문성을 강화해줄 뿐 아니라, 대량의 평가 수요에 기계의 효율성으로 대응할 수 있게 해줍니다. LLM 개발의 병목이었던 'human feedback' 과정을 한층 더 정교화, 자동화할 핵심 열쇠가 될 것으로 기대됩니다.

그림 1 CriticGPT vs 인간 주석자 평가 비교



[6] LLM Critics Help Catch LLM Bugs(Nat McAleeseMaja, LLM Critics Help Catch LLM Bugs, 2024)

4. RLHF 고도화 방안

4.1. CriticGPT 기반 RLHF 개선

이제 CriticGPT 를 기존의 RLHF 파이프라인에 어떻게 녹여낼 수 있을지 고민해 볼 차례입니다. OpenAI 는 크게 세 방향을 제시하고 있는데요[7], 하나씩 살펴보겠습니다. 첫째, CriticGPT 로 사람 평가자들의 피드백 품질을 높이는 겁니다. 버그나 오류를 빠뜨리지 않고 구체적으로 지적할 수 있게 되니 말 그대로 'high-quality human feedback'이 가능해집니다. 이는 더욱 우수한 학습 데이터로 이어져 LLM 성능 향상에 기여할 수 있습니다.

둘째, CriticGPT 를 통해 피드백 과정의 병목 현상을 해소하는 것입니다. 사람의 평가 역량에는 한계가 있기 마련인데, CriticGPT 가 이를 확장해줄 수 있습니다. 마치 바둑에서 AI 가 기보를 대신 검토하듯 말이죠. 평가의 양적 확대는 학습의 반복 주기를 단축시켜 개발 속도를 높일 수 있습니다. 셋째, CriticGPT 와 사람의 협업을 통해 RLHF 의 새로운 지평을 열 수 있습니다. 앞서 보셨듯 CriticGPT 는 사람의 평가를 전방위로 보완해줄 수 있는데요, 여기에 사람 고유의 통찰력까지 결합한다면 지금껏 없던 혁신적인 피드백 체계를 구축할 수 있을 겁니다.

요컨대 CriticGPT 는 인간-AI 심비오시스의 모범 사례로서 RLHF 고도화의 핵심 동력이 될 전망입니다. 평가의 정확성과 효율, 확장성을 두루 높여 ChatGPT 로 대표되는 LLM 의 성능을 한 단계 더 업그레이드하는 데 기여할 수 있을 것입니다.

표 1 CriticGPT를 활용한 RLHF 피드백 개선 방안

구분	내용
장점	● 사람보다 뛰어난 오류 탐지 능력 ● 사람과의 협업을 통한 정확도 향상
보완점	● 텍스트 등 다른 분야로의 확장 필요 ● 평가자 역할 정립 및 잠재적 편향 대응 방안 마련
기대 효과	● 고품질 언어 모델 개발 가속화 ● 실생활 적용 가능 영역 확대

[7] Scalable agent alignment via reward modeling:a research direction (DeepMind, 2018)

4.2. 잠재적 문제점과 해결 과제

CriticGPT 의 활용이 RLHF 고도화의 돌파구가 될 수 있을 만큼 전도유망하지만, 함께 짚고 넘어가야 할 문제들도 있습니다. 무엇보다 주의해야 할 점은 CriticGPT 자체의 편향과 오류 가능성인데요, 이는 모델이 학습한 데이터의 한계에서 비롯될 수 있습니다.

가령 CriticGPT 가 사회적 소수자에 대한 차별적 표현을 문제 삼지 않거나, 특정 이념에 경도된 시각을 정상으로 간주한다면 이는 곧바로 LLM 에 전파돼 악영향을 미칠 수 있습니다. 따라서 CriticGPT 의 공정성과 객관성을 담보하기 위한 지표 개발과 모니터링이 필수적일 것입니다.
또 하나 간과하지 말아야 할 부분은 책임 소재 문제입니다. 만약 CriticGPT 의 피드백을 받아 학습한 LLM 이 잘못된 결과를 내놓는다면 누구에게 책임을 물어야 할까요? 이는 인간-AI 협업 과정에서 반드시 짚고 넘어가야 할 숙제인데요, 역할과 책임에 대한 사회적 합의를 명확히 해두지 않으면 후에 큰 혼란을 겪을 수 있습니다. 가장 바람직한 방향은 CriticGPT 를 비롯한 AI 평가 모델을 사람의 의사결정을 돕는 보조 수단으로 명시하고, 최종 책임은 인간에게 있음을 분명히 하는 것이라 하겠습니다.

5. 향후 전망

5.1. 적용 범위 확대 가능성

지금까지 CriticGPT 의 성과와 잠재력에 대해 살펴보았는데요, 향후 적용 범위는 코딩을 훌쩍 뛰어넘을 것으로 보입니다. 법률, 금융, 의료 등 전문 지식이 필요한 영역에서 CriticGPT 와 같은 AI 평가 모델의 역할이 더욱 커질 전망이거든요. 변호사나 의사처럼 오랜 교육과 실무 경험을 요하는 직군에서는 AI 가 단독으로 의사결정을 내리기엔 한계가 있습니다. 그러나 CriticGPT 처럼 해당 분야의 전문 지식을 내재화한 AI 가 사람 전문가와 협업한다면, 개인의 역량을 뛰어넘는 시너지 효과를 기대할 수 있습니다. 의료 분야를 예로 들어볼까요? 경험 많은 의사와 방대한 의학 데이터베이스를 학습한 CriticGPT 가 힘을 합쳐 환자를 진단한다고 가정해 봅시다. AI 가 수많은 사례와 최신 연구 결과를 참조해 여러 가능성을 제시하고, 의사가 개별 환자의 특성을 고려해 최적의 치료 방향을 결정하는 식이죠. 이런 식으로 의사 개인의 역량을 보완하면서도 숙련된 전문가의 노하우를 담보할 수 있다면, 진단의 정확도와 치료 효과를 획기적으로 높일 수 있을 것입니다. 법률·금융 등 다른 전문 영역에서도 마찬가지겠죠. CriticGPT 가 선도하는 Human-AI 협업은 우리 사회 전반의 지적 역량을 한 단계 업그레이드할 잠재력을 품고 있습니다.

5.2. 연구 개발 방향성 제언

그렇다면 CriticGPT 가 가진 잠재력을 온전히 실현하기 위해 어떤 노력이 필요할까요? 크게 기술적 차원과 사회·제도적 차원의 두 갈래 과제가 있어 보입니다. 우선 기술 개발의 관점에서는 객관적인 성능 평가와 robustness 강화가 시급해 보입니다. 실제 RLHF 파이프라인에서 CriticGPT 를 장기간 운영하며 효과성과 안정성을 실증적으로 검증하는 작업이 필요한데요, 특히 앞서 언급한 잠재적 편향성 같은 리스크 요인을 철저히 분석하고 보완해 나가야 할 것입니다. 동시에 평가 품질과 확장성을 높이기 위해 few-shot learning, 자동화된 하이퍼파라미터 튜닝 등 최신 ML 기법들을 적극 도입하는 것도 중요합니다. 범용적인 평가 모델을 넘어 분야별 전문 평가 모델을 개발하는 것도 검토해볼 만합니다. 기술의 완성도를 높임과 동시에 폭넓은 활용이 가능하도록 업그레이드해 나가야겠죠.

사회·제도적 과제로는 인간-AI 협업의 윤리 규범과 거버넌스 체계 확립을 들 수 있습니다. CriticGPT 와 같은 평가 모델이 법적으로 어떤 지위를 갖고, 그 책임은 어디까지인지 사회적 합의를 모아나가는 게 급선무입니다. EU 의 AI Act 같은 선제적 입법 노력도 좋은 참고 사례가 될 수 있겠네요[8].

또한 AI 평가 과정의 공정성과 투명성을 담보할 수 있는 감독·통제 장치도 마련돼야 합니다. 블랙박스로 남아서는 안 되겠죠. 이를 토대로 사회 전반의 AI 리터러시를 높여 나간다면 CriticGPT 로 대표되는 Human-AI 협업 모델이 보다 건강하게 뿌리내릴 수 있을 것입니다.

6. 결론

지금까지 OpenAI 의 CriticGPT 사례를 통해 LLM 개발에서 인간-AI 협업의 가능성과 과제에 대해 살펴보았습니다. CriticGPT 는 코드 평가에서 사람의 역량을 뛰어넘는 성능을 보여줌으로써, RLHF 고도화를 위한 Human-AI 심비오시스의 새로운 지평을 열어젖혔습니다. 그러나 동시에 편향성 문제나 책임 소재의 불명확성 같은 숙제도 안겨주었는데요, 기술 고도화와 더불어 윤리·사회적 차원의 진지한 논의와 합의 도출 노력이 뒷받침되어야 할 것으로 보입니다. 그래야만 CriticGPT 로

[8] EU Artificial Intelligence Act: The European Approach to AI (KopMauritz, 2021)

대표되는 평가 모델들이 LLM 은 물론, 인간 사회의 역량 향상에 진정으로 기여할 수 있지 않을까요?

AGI(Artificial General Intelligence)의 등장이 가시권에 들어온 지금, CriticGPT 의 사례는 우리에게 시사하는 바가 큽니다. 단순히 기술 발전을 넘어, 인간과 AI 가 조화롭게 공존하며 시너지를 만들어가는 길. 그것이 바로 우리가 마주한 시대적 과제이자 기회라는 사실을 다시금 일깨워주니까요. 적극적인 소통과 협력으로 Human-AI 공진화의 미래를 현명하게 설계해 나가는 것, 그것이 우리 모두에게 주어진 역할이 아닐까 싶습니다.

#CriticGPT, #크리틱 GPT, #ChatGPT, #챗 GPT, #OpenAI, #오픈에이아이, #RLHF, #강화학습, #ReinforcementLearning, #HumanFeedback, #휴먼피드백, #인간피드백, #Alignment, #얼라인먼트, #정렬, #LargeLanguageModel, #대규모언어모델, #LLM, #CodingAssistant, #코딩어시스턴트, #CodeEvaluation, #코드평가, #BugDetection, #버그탐지, #HumanAICollaboration, #인간 AI 협업, #Symbiosis, #공생, #심비오시스, #AugmentedIntelligence, #증강지능, #Ethics, #윤리, #책임, #Responsibility, #편향, #Bias, #Governance, #거버넌스, #AGI, #ArtificialGeneralIntelligence, #범용인공지능

참조 자료

Finding GPT-4's mistakes with GPT-4(Nat McAleeseMaja, 2024)

intended to help identify hallucinations as models grow more sophisticated((STRICKLANDELIZA, 2024)