Abstract

본 논문에서는 다음과 같은 대화형 LLM을 정량적으로 평가하기 위한 프레임워크를 제안합니다.

공개적으로 사용 가능한 데이터 세트를 사용하는 ChatGPT.

우리는 광범위한 기술 평가를 수행합니다.

8가지 일반적인 NLP 애플리케이션 작업을 다루는 23개 데이터 세트를 사용하는 ChatGPT. 우리

이러한 데이터를 기반으로 ChatGPT의 다중 작업, 다중 언어 및 다중 모드 측면을 평가합니다.

세트와 새로 설계된 다중 모드 데이터 세트.

우리는 ChatGPT가 LLM보다 뛰어난 성능을 발휘한다는 사실을 발견했습니다.

대부분의 작업에 대한 제로샷 학습이 가능하며 일부 작업에서는 미세 조정된 모델보다 성능이 뛰어납니다. 우리

비라틴어를 이해하는 것이 더 낫다는 것을 알게 되었습니다.

스크립트 언어를 생성하는 것보다. 그것은

중간 코드 생성 단계를 통해 텍스트 프롬프트에서 다중 모드 콘텐츠를 생성할 수 있습니다. 게다가 우리는 ChatGPT를 발견했습니다.

10가지 항목의 평균 정확도는 63.41%입니다.

논리적 추론 아래의 추론 범주,

비텍스트 추론, 상식 추론으로 인해 신뢰할 수 없는 추론이 됩니다.

예를 들어, 귀납적 추론보다 연역적 추론에 더 좋습니다. ChatGPT는 다른 LLM과 마찬가지로 환각 문제를 겪고 있으며

더 많은 외부 환각을 생성합니다.

외부 지식 베이스에 접근할 수 없기 때문에 파라메트릭 메모리입니다. 마지막으로,

ChatGPT의 대화형 기능을 통해 기본 LLM과의 인간 공동 작업이 가능합니다.

즉, 요약에서 8% ROUGE1, 기계 번역에서 2% ChrF++를 여러 차례의 "프롬프트 엔지니어링" 방식으로 개선합니다. 우리는 또한

평가 세트 추출.

<https://github.com/HLTCHKUST/chatgpt-evaluation/blob/main/main.py>

1.troduction

ChatGPT는 대규모 언어의 후속 제품입니다.

모델(LLM) InstructGPT(Ouyang 외, 2022)

다음을 사용하여 미세 조정된 대화 상자 인터페이스를 사용합니다.

인간 피드백을 통한 강화 학습

1

https://github.com/HLTCHKUST/chatgpt-evaluat

이온

(RLHF) (Christiano et al., 2017) 접근법.2

에서

지난 몇 달 동안 ChatGPT가 모였습니다

사용자 수는 100만 명(Hu, 2023)으로 증가하고 있으며,

기업과 소비자 모두에 의해 수많은

대부분 텍스트 작업입니다. 전례 없는 인기를 누리는 이유 중 하나는 ChatGPT가 그 규모를 통해

RLHF를 통해 인상적인 능력을 보여주었습니다.

NLP의 다양한 영역과 창발적 능력

코드 생성, 멀티모달 생성 등이 있습니다. 또 다른 이유는 대화 인터페이스 때문입니다.

사용자가 기본 대형과 상호 작용할 수 있습니다.

언어 모델을 통해 보다 효과적이고 효율적으로

다중 턴 프롬프트와 유사한 대화형 채팅

공학.

그러나 강력한 기능에도 불구하고 ChatGPT에 대한 일화 보고서는 지속적으로

중요한 남은 과제 - 예를 들어,

일부 초등 수학에서는 실패합니다(Gilson

외, 2022; 골드버그, 2023; 프리더 외, 2023;

최 외, 2023; Davis, 2023) 및 상식

추론 작업(Guo et al., 2023; Davis, 2023);

진실에 근거하지 않은 것에 대해 인간과 같은 유창함과 웅변으로 환각을 일으킨다.

외, 2023; 소프, 2023; 스미스, 2023); 그리고

다음에서 훈련된 범용 언어 모델

웹에 있는 모든 것의 언어 범위는

의문스럽다(Lu et al., 2022; Jiao et al., 2023).

OpenAI는 ChatGPT의 많은 제한 사항을 나열했습니다.

3 CEO는 “실수입니다.

중요한 일을 [ChatGPT]에 의존하는 것

지금 당장”(Altman, 2022). 많은 연구자들은

겉모습에도 불구하고 ChatGPT와 같은 LLM은 언어 능력만 뛰어나고 실제로는 그렇지 않다고 주장했습니다.

추론(Mahowald et al., 2023). Consequently 전

결과적으로 사람들이 무엇을 할 수 있는지 또는 무엇을 할 수 있는지는 명확하지 않습니다.

인기에도 불구하고 사용할 수 없습니다. 사용자용

연구자들 모두에게 다음과 같은 정보를 제공하는 것이 도움이 될 것입니다.

https://beta.openai.com/docs/model-index-for

-연구원

https://platform.openai.com/docs/chatgpt-edu

양이온

arXiv:2302.04023v2 [cs.CL] 2023년 2월 28일

다양한 분야에서의 신뢰성에 대한 자신감

NLP/AI 작업.

이전 연구에서는 ChatGPT와 관련된 윤리적 영향이나 우려 사항을 논의했습니다.

기타 LLM) (Jabotinsky 및 Sarel, 2022; Susnjak, 2022; Blanco-Gonzalez 외, 2022; Aydın

Karaarslan, 2022; Jeblick 외, 2022). 하지만 기술적인 평가는 많이 이루어지지 않았습니다.

ChatGPT4의 강점과 한계

. 에게

이 격차를 메우기 위해 ChatGPT에서 실험을 수행합니다.

질문 답변, 추론, 요약, 기계 번역, 자동 번역 등 주요 NLP 작업에 대한 표준 공개 테스트 세트의 샘플 포함

사후 편집, 감성 분석, 언어 식별, 작업 중심 대화(대화 상태)

추적 및 응답 생성) 및 잘못된 정보 탐지. 다국어 성능과 비전-언어 다중 모드를 평가합니다.

능력. 추가 실험을 통해 우리는 또한

주요 한계를 정량적으로 평가합니다.

추론과 환각. 또한 다중 회전 상호 작용을 테스트하기 위한 실험을 수행합니다.

더 나은 신속한 엔지니어링을 위한 수단으로. 우리는 바랍니다

ChatGPT 사용자에게 통찰력을 제공하기 위해

위에서 언급한 장점과 한계 또한

상호작용을 통해 어떻게 결과를 개선할 수 있는지 알아보겠습니다. (정량적으로 측정할 수는 없습니다.

사용자 로그에 액세스하지 않고도 ChatGPT의 RLHF 측면을 평가합니다. OpenAI가 출시되기를 바랍니다.

이 작업을 수행하면 그러한 평가를 수행할 수 있습니다.

앞으로는 OpenAI와 협력해 나갈 예정입니다.)

다음은 우리가 가지고 있는 주요 통찰력입니다.

평가를 통해 얻은 것:

-Multitask, Multimodal, and Multilingual

9/13 NLP 데이터 세트의 경우 ChatGPT는 제로샷 학습으로 이전 LLM보다 성능이 뛰어납니다. 심지어 4가지 다른 작업에서 완전히 미세 조정된 작업별 LM보다 성능이 뛰어납니다. 다른 곳에서는

경우 ChatGPT는 동등하거나 약간 낮습니다.

특정 NLP 작업에 맞게 완전히 미세 조정되는 것보다

• ChatGPT가 낮은 리소스로 일반화되지 않음

리소스가 매우 적은 언어(예:

마라티어, 순다어, 부기니어). 거기

리소스가 부족한 언어, 특히 라틴어가 아닌 언어의 전반적인 성능 저하입니다.

번역의 경우 스크립트; 약점

이해보다는 세대에 있다

번역 과정의 일부;

4많은 일화적인 분석이 온라인에 게시되었지만

포괄적으로 없음

• ChatGPT는 코드 중간 매체를 가능하게 합니다.

시각과 언어를 연결하는 역할을 하지만

다중 양식 능력은 아직 초보적입니다.

비전 언어 모델과 비교.

추론 우리는 총 634개의 샘플을 사용하여 10가지 추론 범주를 테스트했습니다. 우리의 기준으로

실험에서 ChatGPT는 더 많은 약점을 보여줍니다.

연역적 또는 귀추적 추론보다 귀납적 추론

추리. ChatGPT에는 공간 추론도 부족합니다.

더 나은 시간적 추론을 보여 주면서. 채팅GPT

또한 수학적 추론이 부족합니다.

Frieder 등의 최근 연구 결과를 바탕으로. 또한, 우리는

ChatGPT는 비텍스트 의미 추론보다 상식 추론에 상대적으로 더 뛰어나다는 사실을 발견했습니다. 마지막으로 ChatGPT는 인과 및 유추 추론에서 수용 가능한 성능을 보여주지만

다음과 유사한 다중 홉 추론 기능이 좋지 않습니다.

복잡한 추론에 있어서 다른 LLM의 약점(Ott

외, 2023).

른 LLM과 유사한 환각(Radford <Hallucination>

외., 2019; Muennighoff 외, 2022; 작업장

et al., 2022), ChatGPT는 환각 문제를 겪고 있습니다. 이는 더 많은 외적 환각, 즉 매개변수 메모리에서 소스로부터 확인할 수 없는 사실 진술을 생성합니다.

액세스 권한이 없기 때문에 모든 작업에 걸쳐

외부 지식 기반에.

상호작용성 주요 차별화 요소 중 하나

이전 버전의 ChatGPT 요소는

다중 회전 대화 상호 작용. 이를 통해 ChatGPT는 대화 세션 내에서 여러 작업을 수행할 수 있습니다. 또한 성능이 크게 향상되었습니다(요약 시 ROUGE-1 8% 및

저자원 기계 번역 시 ChrF++ 2%)

다양한 표준 NLP의 다중 회전 상호작용을 통해

작업. 이 프로세스는 프롬프트 엔지니어링과 유사합니다.

시스템의 피드백을 통해

본 논문의 구성: 우리는 먼저

ChatGPT 및 관련 작업 개요(§2)

이후 ChatGPT를 통해 평가 결과를 제공해드립니다.

다양한 애플리케이션 테스트 세트, 다국어 테스트

세트 및 §3의 새로운 다중 모드 작업에 대해 설명합니다. 그러면 우리는

세 가지 주요 강점과 약점을 살펴보세요.

ChatGPT, 즉 추론(§4), 환각

(§5) 및 후속 세 가지의 상호작용성(§6)

섹션. 마지막으로 논의하고 결론을 내린다.

ChatGPT 결과에 대해 알아보세요.

4~5 reasoning hallucination