Factuality challenges in the era of large language models and opportunities for fact-checking

Isabelle Augenstein[1], Timothy Baldwin[2], Meeyoung Cha[3], Tanmoy Chakraborty[4] ✉, Giovanni Luca Ciampaglia[5], David Corney[6], Renee DiResta[7], Emilio Ferrara[8], Scott Hale[9], Alon Halevy[10], Eduard Hovy[11], Heng Ji[12], Filippo Menczer[13], Ruben Miguez[14], Preslav Nakov[2], Dietram Scheufele[15], Shivam Sharma[4] & Giovanni Zagni[16]

연구 배경

- LLM 기반 도구가 자연어 생성 능력에서 높은 성과를 보였다
- LLM 기반 도구는 잘못된 정보나 오류를 생성할 가능성이 크다
- LLM 기반 도구는 악용되어 사회적 문제를 일으킬 수 있다
- Fact Checking의 중요성이 부각되고 있다

LLM의 등장

- 1948년 Claude Shannon이 인간의 언어와 유사한 텍스트를 생성할 수 있는 통계 언어 모델을 제안
- 계산의 발전 덕분에 방대한 양의 텍스트를 학습할 수 있게 됨
- 사람의 텍스트와 구별하기 어려운 텍스트 생성 가능
- 자연어에 대한 기억과 추론에 있어 좋은 성능 달성

LLM과 LLM 기반 챗봇의 위험성

- Faithfulness: 생성된 텍스트가 입력 컨텍스트에 충실한가
- Factuality: 생성된 텍스트가 사실인가
  - LLM은 최신 데이터가 부족함
    - 코로나 19 팬데믹 기간 동안 6594건 중 30%가 코로나19라는 키워드를 사용해 문의
  - LLM이 검색 도구에 적용됨으로 인해 온라인 생태계 건전성에 대한 우려
    - 신뢰할 수 있던 검색 도구였지만 LLM으로 인해 신뢰성이 깨짐
    - Bard: 제임스 웹 망원경의 외계 행성 이미지 발견을 거짓이라고 주장
    - Replika: 빌 게이츠가 코로나19의 설계자임을 암시
  - 거짓이거나 오해의 소지가 있는 콘텐츠를 대규모로 생성할 수 있음
    - 2020년 미국 대선 기간 동안 선거 사기에 대한 허위 기사 생성
    - LLM은 자신이 모르는 것이 무엇인지 모름

기존 연구

- LLM의 사실성에 대한 연구
  - LLM이 콘텐츠를 어떻게 저장, 처리, 생성하는지 설명
  - 특정 영역에서 문제를 분류하고 개선 전략을 평가
- LLM의 환각에 대한 연구
  - 내재적 및 외재적 환각, 출처 및 완화 방법을 식별
  - 환각을 입력 충돌형, 맥락 충돌형, 사실 충돌형으로 분류
  - 출처, 평가 조치 및 완화 전략을 탐색
  - 의료 및 법률과 같은 영역의 윤리적 영향을 설명

- But, LLM의 역할에 대한 구체적인 설명이 없음

LLM의 사실성 문제(1)

- Citation Gaps
  - LLM은 신뢰할 수 있는 자원이 부족
  - 생성된 문장의 절반이 인용문이 부족하고 자신의 주장을 뒷받침하는 문장은 4분의 3에 불과

- Grounding Deficiency
  - 모델이 생성한 텍스트가 지식이나 데이터에 대한 충분한 연결이 없음
  - 내용이 사실에 기반하기 않기 때문에 신뢰할 수 없음

LLM의 사실성 문제(2)

- Truthfulness
  - LLM이 여전히 Hallucination 생성
    - Factual Inaccuracy: 사실적으로 정확한 텍스트 제작 어려움
    - Contextual coherence: 문맥적으로 맞지 않는 응답 생성 가능
    - Domain-specific reliability: 성능이 도메인에 따라 크게 달라짐
    - Deductive VS Inductive reasoning: 귀납적 추론보다 연역적 추론에 능숙함

- Confident Tone
  - 정확성이 손상되더라도 자신감 있는 어조를 유지
  - 사실성에 대한 측정과 관련 불확실성을 표현할 수 있는 방법이 부족
  - Do-Not-Answer 같은 데이터 셋이 있지만 쉽게 우회 가능

LLM의 사실성 문제(3)

- Fluent style
  - LLM의 유창한 글쓰기 스타일은 설득력을 높임
  - 신뢰할 수 없는 출처의 주장에 대해 모순된 지식에도 불구하고 진실에 대한 인식을 형성하게 함

- Direct use
  - 사용자와 ChatBot의 상호작용 과정에서 유효성 식별 어려움
  - ChatBot이 전달하는 잘못된 정보가 다른 많은 사실 및 거짓 주장에 묻혀 탐지 및 수정 어려움

LLM의 사실성 문제(4)

- Halo Effect
  - 일부에 대한 능숙함이 모델 전체의 성능에 긍정적인 영향을 끼침
  - 단일 특성이 사용자의 인식에 크게 영향을 미쳐 잘못된 평가로 이어질 수 있음

- Unreliable evaluation
  - 벤치마크 데이터 셋만으로는 충분하지 않음
  - 훈련 중 오염되었을 수 있음
  - GPTScore, G-Eval, SelfCheckGPT
  - Ground-Truth 참조 유지는 비용이 많이 들고 Data drift의 영향을 받음

LLM의 사실성 문제(5)

- Outdated Knowledge
  - 과거 데이터에 의존하면 LLM의 효율성이 낮아짐
  - LLM 업데이트 과제
    - 지식 편집의 파급 효과 평가
    - 관련 없는 사실에 영향을 미치지 않고 관련된 사실만을 업데이트 하도록 보장
    - 오래된 지식 업데이트에 대한 저항
    - 여러 언어에 걸쳐 일반화 어려움
  - LLM은 어텐션에 중점을 둔 연관성에 의존
  - 학습 데이터 불균형
  - 사전 훈련된 지식과 새로운 지식 통합 어려움

악의적인 LLM 사용으로 인한 위협(1)

- Personalized attack
  - 사용자의 과거 대화/글의 로그를 프롬프트에 통합해 허위 정보, 피싱 메시지 생성 가능
  - 사실과 혼합하여 설득력 높일 수 있음
  - 오픈 LLM 조작하여 개인 정보 추출해서 피싱 공격 및 사기에 사용 가능

- Style impersonation
  - 특정 인물의 글쓰기 스타일을 모방한 텍스트 생성 가능
  - 인물의 신뢰도를 떨어뜨리기 위해 소셜 미디어 플랫폼에 배포 가능

악의적인 LLM 사용으로 인한 위협(2)

- Bypassing detection
  - “동일한 콘텐츠에 반복적으로 노출되는 경우는 드물다”
  - LLM은 동일한 주장에 대한 무한 변형 생성 가능
  - 변형된 콘텐츠의 누적 영향은 Fact Checker에게 보이지 않음

- Fake profiles
  - LLM을 통해 신뢰할 수 있는 가짜 프로필과 콘텐츠 생성 가능
  - Bad Actor가 온라인 커뮤니티에 쉽게 침투 가능
  - 불법이거나 조작적인 콘텐츠를 퍼뜨려 왜곡된 인식 만들 수 있음
  - 생명을 위협하는 문제로 확대 가능

위협 해결(1)

- Alignment and safety
  - 데이터 정리, 피드백을 통한 강화 학습, 프롬프트 데이터 셋 큐레이션 및 머신 러닝 기반 필터링 같은 조치를 통해 해결되는 주요 관심사
  - Red Teamming 과정에서 응답의 사실성에 미치는 영향에 대한 연구 부족
  - 사실성 인식: 환각을 줄이기 위해 파인 튜닝 과정 감독
  - 사실성 및 지침 준수를 위한 별도의 보상 모델을 통해 직접 선호도 최적화
  - 오래되었거나 잘못된 정보를 수정하기 위한 지침 재정의
  - 사실성-선호도 쌍 및 검색 증강 클레임 확인과 같은 대응책
  - 많은 오픈 소스 모델에 아직 안전장치 부족
  - 규제는 명확하지 않지만 LLM의 부정적 영향을 완화하는 것 필수
  - LLM의 부작용을 줄이기 위해 가능한 모든 조치를 추구해야 함

위협 해결(2)

- Retrieval-augmented generation(RAG)
  - 외부 소스의 컨텍스트 정보를 텍스트 생성에 통합
  - 부정확한 콘텐츠 생성 문제 완화
  - 응답 품질 개선, 데이터 민감도 감소, 훈련 비용 감소, 복잡하고 긴 텍스트 쿼리 처리 개선 등 LLM의 주요 한계 완화할 잠재력

- Hallucination mitigation
  - 최근의 시도는 자기 일관성 확인, 교차 모델 검증 또는 RAG를 기반으로 추론 단계에서 LLM 환각 문제를 해결하는 데 중점을 둠
  - 사전 교육, 파인 튜닝 감독, 피드백을 통한 강화 학습 및 추론 등 다양한 시스템 모델링 단계에서 연구

위협 해결(3)

- Knowledge updating and maintenance
  - 지식 편집 활성화하면 모델 직접 업데이트해 사실 오류 수정 가능
    - 편집의 파급 효과를 평가하여 관련 내용만 업데이트 되도록 하는 것
  - 잘못된 정보를 생성하고, 오래된 지식의 업데이트를 거부
  - 훈련 과정에서 단어 간 연관성을 통해 지식을 학습하기 때문에 다국어 일반화 어려움
  - 비직관적인 업데이트 방식이 문제를 심화시키며 새로운 지식이 반영되지 못하게 됨
  - Liu의 새로운 프레임워크는 훈련 없는 자체 대조 디코딩 접근 방식을 사용해 실제 사건을 기반으로 사실 지식을 업데이트 하는 것 강조

위협 해결(4)

- Alleviating historical training bias
  - 과거의 편향은 고정 데이터 셋에서 훈련된 LLM의 출력 품질에 영향
  - 훈련 데이터를 넘어 효과적으로 일반화할 수 있는 LLM의 능력을 개선해 문제 해결
    - 관련 명령어-응답 쌍을 동적으로 도출해 선택적 업그레이드

- Better evaluation
  - BERTScore 같은 기존 평가 지표는 LLM의 진화하는 기능 및 요구사항에 맞지 않음
  - GPTScore 및 G-Eval은 다양한 작업에서 합리적인 상관관계를 보임
    - but 개선의 여지 많음
  - 잠재적 방향성: 특정 영역에 대한 사실성 지침 맞춤화

위협 해결(5)

- Recognizing AI-generated content
  - LLM의 출력과 사람이 작성한 텍스트 구별하기 어려움
  - 가짜 뉴스 탐지기 성능 저조
  - 워터마크 우회 쉬움
  - 서로 다른 탐지기를 사용해 여러 언어에 대해 여러 도메인에서 여러 생성기를 연구해야 함
  - 최신 기계 생성 콘텐츠 컬렉션을 유지해야 함
  - 강력하고 완벽한 탐지 기술 고안해야 함

위협 해결(6)

- Content authenticity and provenance
  - 사람이 작성한 텍스트의 가치가 높아질 수 있음
  - 많은 사람에게 도달하기 전에 가짜 콘텐츠의 확산을 제한해야 함
  - 콘텐츠 진위 여부 및 입증 기술은 비디오 및 이미지 콘텐츠에 존재
    - 콘텐츠 생성 관련 메타데이터가 변경되지 않음을 증명할 수 있는 암호화 서명 방법
- 책임감 있는 공개 및 개방형 연구
- 조정 및 협업
- 규제
- AI 문해력 증진
- 기술 개발

Fact Checking의 전망(1)

- 쿼리 표현, 컨텍스트 정보 검색, 설명 가능성 및 추론, 다국어 기능, 요약 등 LLM 기능의 이점을 누릴 수 있음
- 클레임 검증에 사용 가능
- 연설, 토론, 뉴스 방송 전사, 광범위한 문서 요약, 클레임 목록 간결화
- 이전에 사실 확인된 클레임을 반복하거나 의미가 동일한 클레임 식별
- 주의 사항
  - 긴 문서를 요약할 때 오류의 위험
  - 인간이 AI 지원 Fact-Checking과 어떻게 상호 작용하는지 이해해야 함

Fact Checking의 전망(2)

- Stance Detection
  - LLM이 Fact Check 과정에서 보조 역할을 할 수 있음
  - 주장의 진실성을 결정하지 않음 but Fact Checking 프로세스에 유용
  - LLM이 테스트 데이터를 확인했을 수 있으며 특정 영역에서 성능이 낮을 수 있음
  - ChatGPT는 감정이나 실용적 분석과 관련된 작업보다 Stance Detection에서 더 우수함
- Domain-specific Verification
  - 도메인별로 검증된 정보 검색
  - 여전히 잘못된 정보 삽입 가능

시사점

- 사실 검증의 중요성
- LLM의 사실성 문제
- LLM의 역할에 초점을 맞춰 연구 진행
- LLM을 활용한 Fact-checking의 기회 제시

소감

- 논문에서 LLM의 단점과 그로 인해 발생할 수 있는 위협, 그에 대한 해결책을 체계적으로 다루며, LLM의 미래 방향성을 잘 제시한 것 같다
- 무의식적인 LLM 사용을 조심해야 한다고 느낌
- 불완전한 LLM도 Fact Checking에 도움이 될 수 있음을 알게 됨

Q&A

Open Question

- 논문에서 다룬 내용 이외의 LLM의 단점은 과연 더 없을까?
- Deepfake