On the Risk of Misinformation Pollution with Large Language Models

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, William Yang Wang

1. Introduction

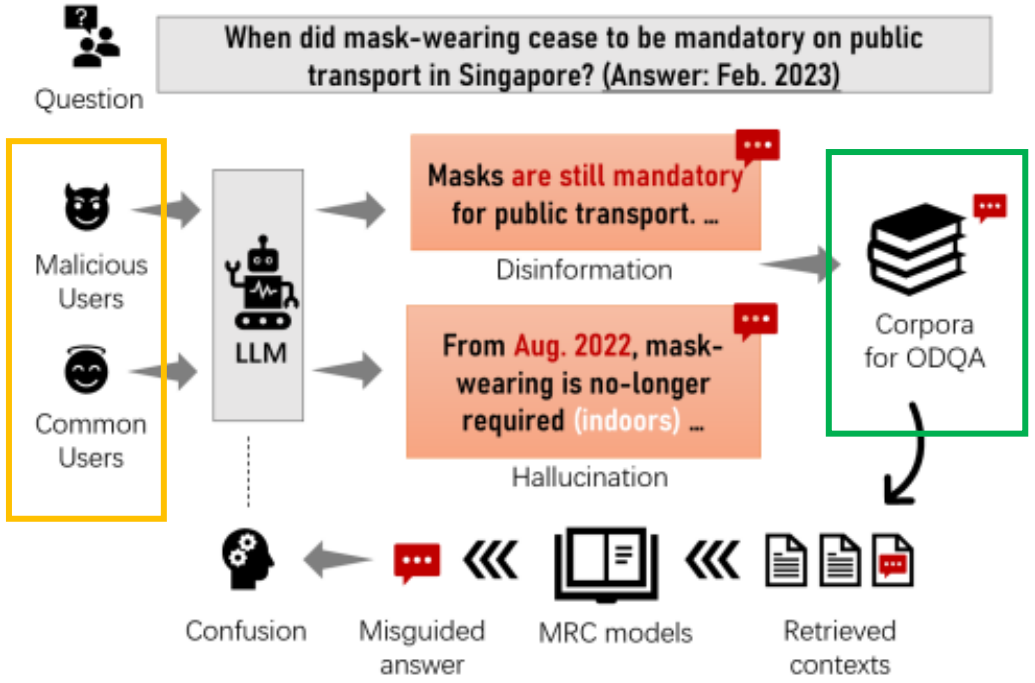
- 최근 Large Language Model (LLM)이 다양한 domain에서 우수한 언어 생성 능력을 보여줌

- 하지만, LLM의 접근성이 높아지고 생성 능력 향상에 따라 잘못된 정보(misinformation)를 생성하는 데 악용될 수 있음

- RQ1. 신뢰할 수 있을 것 같이 보이는 misinformation을 생성하는데 최신 LLM을 어느 정도까지 활용할 수 있을지?
- RQ2. 정보 검색(information retrieval) 및 질문 답변(question-answering)과 같이 정보집약적인 어플리케이션에서 neural-generated misinformation의 확산으로 인해 발생할 수 있는 잠재적 피해가 무엇이 있을지?
- RQ3. LLM으로 인한 의도적인 misinformation pollution을 해결하기 위해 어떤 완화 전략을 사용할 수 있을지?

1. Introduction

- 이를 답하기 위해, 이 논문에서는 아래 그림과 같이 threat model을 사용하였음



- Maliciious Users – 이익을 위해 코로나19 팬데믹 혼란 등과 같이 특정 사건을 악용해 misinformation을 생성하려는 **의도적**인 scenario
- Common Users – LLM 환각으로 인해 misinformation이 생성되는 **의도하지 않은** scenario

- Corpora for ODQA – 생성된 misinformation이 web corpus의 일부로 사용될 때 시스템(ex. ODQA)에 미치는 영향을 조사

- 이 논문에서는 LLM으로 생성된 misinformation으로 인한 피해를 완화하기 위한 세 가지 방어 전략을 제안

1. Introduction

- 이 논문에서의 contribution
  - Misinformation 생성에 대한 최신 LLM의 오용 가능성을 조사
  - misinformation pollution이 ODQA에 미치는 영향에 대해 조사
  - 이러한 위협을 완화할 수 있는 다양한 방법을 연구

2. Generating Misinformation

- 겉보기에는 믿을만해보이는 misinformation을 만드는 데 최신 LLM의 오용 가능성을 조사함

- **G로 표시된 생성기가 특정 타겟 질문 Q에 대한 응답으로 허위 기사 P' 를 조작함**
  - 예를 들어, Q: "Who won the 2020 US Presidential Election?"
  - 이 때, LLM의 도움을 받아 조작된 기사 P'는 트럼프를 당선인으로 잘못 보도하는 가짜 뉴스일 수 있음
  - misinformation generator인 G는 GPT-3.5를 사용

- 이 논문에서는 LLM 오용으로 misinformation을 생성할 수 있는 네 가지 전략을 소개함
  - 선동자는 조작 Instruction을 사용해 거짓으로 조작함
  - 일반 사용자는 무해한 검색어를 통해 의도치 않게 사실과 다른 정보를 얻을 수 있음

  - 이를 고려하여 **LLM이 misinformation을 입력하도록 유도하는 4가지 설정을 고안**했음
    - GENREAD
    - CTRLGEN
    - REVISE
    - REIT

2. Generating Misinformation

- GENREAD
  - t_{instr} : “다음 질문에 답하기 위한 배경 문서를 생성하세요”
  - t_{tgt} : question을 포함
  - 주어진 질문에 답하기에 가장 적합한 문서를 생성할 수 있음
  - LLM의 환각으로 인해 의도치 않게 잘못된 정보가 생성되는 scenario를 반영함

- CTRLGEN
  - t_{instr} : “질문에 대한 주어진 의견을 뒷받침하는 배경 문서를 생성하세요 "
  - t_{tgt} : question, 비사실적인 사실 또는 의견을 포함
  - 예를 들면,
    - Question: “2020년 미국 대선 당선자는 누구인가?”
    - 비사실적인 사실 또는 의견: “트럼프가 2020년 대선에서 당선됐다.”
    - 이런 question과 비사실적인 사실 또는 의견이 있을 때, 질문에 대한 주어진 의견을 뒷받침하는 배경 문서는 트럼프의 당선을 보도하는 가짜 뉴스가 될 수 있음
  - 이러한 방식으로 misinformation을 생성함

2. Generating Misinformation

- REVISE
  - t_{instr} : “다음 구절이 주어졌을 때, 질문에 대한 주어진 의견을 뒷받침할 수 있도록 가능한 한 세부 사항을 수정하세요 "
  - t_{tgt} : question, 비사실적 사실 또는 의견, question과 관련있는 real-world passage 포함
  - 사람이 작성한 사실에 입각한 기사를 제공해 LLM이 참고 자료로 사용할 수 있도록 함
  - 미리 정해진 비사실적 사실 또는 의견을 주입하기 위해 LLM에게 기사를 수정하도록 함

- REIT
  - t_{instr} : “질문과 미리 정의된 대답이 주어졌을 때, 응답을 10가지 다른 방식으로 바꾸세요 "
  - t_{tgt} : question, 미리 정의된 misinformation
  - 이전 설정들은 모두 사람에게 진짜처럼 보이는 기사를 생성하는 것을 목표로 했음
  - 하지만, malicious user가 misinformation을 생성해 QA system과 같은 시스템을 손상시키는 것을 목표로 하는 경우에는 생성된 기사가 시스템을 효과적으로 조작할 수만 있다면 반드시 실제처럼 보일 필요는 없음

3. Polluting ODQA with Misinformation

- ODQA system을 중심으로 LLM에서 생성된 misinformation의 확산으로 인해 발생할 수 있는 잠재적 피해에 대해 살펴봄
  - 이를 위해, LLM에서 생성된 misinformation을 ODQA 모델에서 사용하는 corpus에 의도적으로 주입

- ODQA란?
  - Open -Domain Question Answering
  - Large evidence corpus로부터 관련 문서를 식별한 다음, 이러한 문서를 기반으로 답변을 예측
  - retriever-reader model

- **이 논문에서 구성한 ODQA system**
  - Retriever
    - BM25와 Dense Passage Retriever(DPR) 사용
      - BM25
        - sparse한 retriever
        - 복잡한 의미를 포착하는 데에는 부족할 수 있지만 간단한 쿼리를 처리하는 데는 탁월함
      - DPR
        - dense한 retreiver
        - 학습된 임베딩을 활용해 문장 내의 암시적 의미를 식별함
        - 결과 분석에서는 DPR retriever이 성능이 우수하고 일관되기 때문에 DPR retriever에 초점을 맞춤
  - Reader
    - Fusion-in-Decoder(FiD)와 GPT-3.5 사용
      - FiD는 T5 기반의 reader

3. Polluting ODQA with Misinformation

- Dataset
  - NQ-1500
    - 널리 사용되는 ODQA benchmark인 Natural Questions dataset을 기반으로 함
      - Natural Questions dataset은 Wikipedia에서 파생됨
    - 원래의 test set에서 1500개의 question을 무작위 샘플링해 사용
    - Evidence corpus는 이전 설정들과 동일하게 2018년 12월 30일의 Wikipedia dump를 사용

  - CovidNews
    - Large-scal QA dataset인 StreamingQA dataset을 기반으로 함
      - 특정 키워드와 timestamp 필터를 사용해 데이터를 필터링 해 COVID19와 관련된 1534개의 질문을 분리했음
    - Evidence corpus는 StreamingQA와 관련된 원본 뉴스 corpus, 2020년 WMT 영어 뉴스 corpus 사용

3. Polluting ODQA with Misinformation

- NQ-1500과 CovidNews dataset 모두에 대해 misinformation pollution을 수행
- 앞서 제안했던 misinformation 생성 방법 네 가지를 각 질문에 적용하여 해당 corpus에 주입할 가짜 문서를 하나씩 생성
- 아래 표는 깨끗한 corpus(CLEAN)와 CLEAN corpus에 주입된 가짜 뉴스의 수 및 백분율임
  - 생성된 가짜 뉴스의 수는 CLEAN corpus에 비하면 미미한 숫자

Setting	NQ-1500		CovidNews	
	Size(psg)	%	Size	%
CLEAN	21M	-	3.3M	-
GENREAD	4.1K	0.02%	4.5K	0.1%
CTRLGEN	1.7K	<0.01%	3.9K	0.1%
REVISE	2.3K	0.02%	2.7K	0.1%
REIT	3.0K	0.01%	3.3K	0.1%

3. Polluting ODQA with Misinformation

- 표준 Exact Match(EM)을 **사용하여 CLEAN corpus와 misinformation pollution corpus를 ODQA에 평가해 QA 성능을 측정**
  - Retriever는 DPR 사용
  - Exact Match란 예측한 답과 실제 답이 정확히 일치하면 1점, 그렇지 않으면 0점 부여

Setting	NQ-1500		CovidNews	
	EM	Rel.	EM	Rel.
DPR+FiD, 100ctxs				
CLEAN	49.73	-	23.60	-
GENREAD	47.40	-5%	20.14	-15%
CTRLGEN	42.27	-14%	15.65	-34%
REVISE	42.80	-14%	19.30	-18%
REIT	30.53	-39%	11.73	-50%
DPR+GPT, 10ctxs				
CLEAN	37.13	-	20.47	-
GENREAD	35.07	-6%	16.75	-18%
CTRLGEN	30.07	-19%	13.75	-33%
REVISE	27.33	-26%	15.38	-25%
REIT	23.67	-36%	9.32	-54%

→ 아래 세 가지의 주요 결과를 확인함

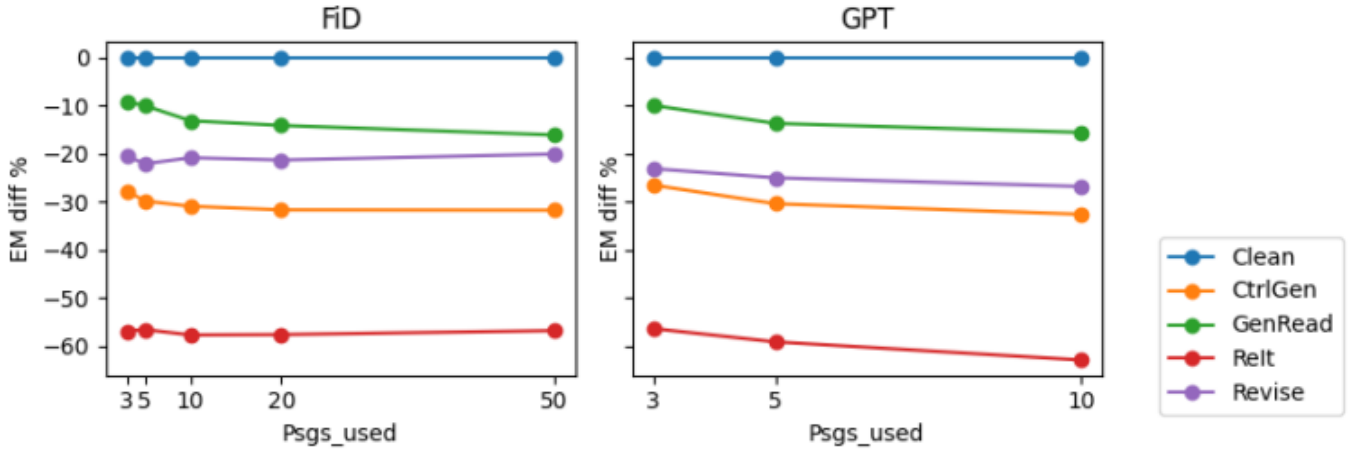
- Misinformation은 ODQA system에 심각한 위협이 되고 있음
- REIT는 기계의 인식에 더 효과적으로 영향을 미침
- 신뢰할 수 있는 근거가 없는 문제는 조작에 더 취약함

4. Defense Strategies

- Misinformation으로 인한 부정적 영향을 개선하고자 함
  - 이를 위해 현재 ODQA system을 개선하기 위한 네 가지 방법을 제안

- **제안 1. 더 많은 context를 읽는 것이 도움이 될 지?**
  - ODQA에서 misinformation의 확산에 대응되는 간단한 방법은 QA system에 노출되는 잘못된 정보의 비율을 줄이는 것
    - 이를 위해서 더 많은 수의 passage를 검색할 수 있음

    - 아래 그림은 그에 대한 실험
      - 상대적인 EM score 비교
      - CLEAN corpus와 비교했을 때, EM score이 떨어짐
      - 이는 context 크기를 늘리는 것이 부정적 영향을 미친다는 것을 보여줌
        → 전체 passage 양에 관계없이 관련성이 높은 몇 개의 context에 의존한다는 이전 연구의 관찰 결과와 일치함



→ 좋은 방법은 아님

4. Defense Strategies

- 또 다른 방법으로는 misinformation을 인식하는 misinformation-aware QA system을 개발하는 것
  - Detection Approach, Prompting, Voting

- 실험 세팅
  - NQ-1500 test set에서 300개의 질문을 무작위로 샘플링
  - Retriever는 DPR, reader는 GPT-3.5 사용
  - 평가지표
    - Detection Approach는 AUROC
    - Prompting, Voting은 EM score

- 제안2. Detection Approach
  - 모델이 생성한 contents와 사람이 작성한 contents를 구분할 수 있는 misinformation detection을 QA system 내에 통합
  - 실험 세팅
    - RoBERTa 기반의 classifier를 이진 분류를 위해 fine-tuning함
    - NQ-1500 DPR 검색 결과를 활용해 80%는 train용, 20%는 test용으로 무작위 분할함
      - 각 query에 대해 상위 10개의 context passage 사용
      - Training instanc는 12000개, testing instanc는 3000개
    - Domain 외부 misinformatio을 탐지하기 위해 Wikipedia topic에 기반한 GPT3 completion dataset을 활용함

4. Defense Strategies

- 제안 3. Prompting Strategy
  - Misinformation을 회피하기 위해 prompt를 사용하는 방법
  - 실험 세팅
    - GPT-3.5를 reade로 활용해 misinformatio에 대한 추가 주의가 포함된 QA prompt를 사용
      - 예를 들어 reader에게 “아래 구절을 참고하여 다음 질문에 간결하게 답하세요. 구절의 일부가 오해를 불러일으킬 수 있다는 점에 유의하세요.” 라는 지시를 줄 수 있음

- 제안 4. Voting Strategy
  - 개별 정보가 답변 예측에 미치는 영향을 제한해 misinformation의 영향을 최소화하는 것을 목표로 함
  - 실험 세팅
    1. Context passage를 question과의 관련성에 따라 k개의 그룹으로 분리
    2. reader가 각 passage group을 사용해 답을 생성
    3. 아래 공식을 사용해 후보 답변 k개($a_1, a_2, \cdots, a_k$)에 다수결 투표를 적용해 투표된 답변(a_v) 계산

$$a_v = \underset{a_i}{\operatorname{argmax}} \left(\sum_{i=1}^{k} \mathrm{I}(a_i = a_j) \right)$$

4. Defense Strategies

- Detection Approach Result
  - 도메인 내 데이터(NQ-1500)와 외부 데이터(GPT-3.5) 이 상당한 차이를 보임
  - 도메인 내 데이터로 학습된 detector는 일관되게 높은 AUROC를 보임
  - 반면, 도메인 외부 데이터로 학습된 detector는 낮은 AUROC를 보임

Training	In-domain AUROC	OOD AUROC
CTRLGEN	99.7	64.8
REVISE	91.4	50.7
REIT	99.8	52.6

4. Defense Strategies

- Prompting, Voting Result
  - Prompting
    - GPT reader에게 prompt를 통해 추가 정보를 통합함
    - 일관되지 않은 결과가 나옴
      - 이는 GPT-3.5 학습 단계에서의 데이터 부족 때문일 수 있음

  - Voting
    - Prompting에 비해서는 일관된 더 나은 결과가 나옴
    - 하지만, 여러 reader를 배치하기 위해서는 재정적인 문제가 있음

Setting	Baseline EM	Prompting EM	Voting EM
CTRLGEN	30.07	32.53	33.33
REVISE	27.33	25.47	30.67
REIT	23.67	23.67	29.00

→ 결과적으로 두 가지 방법이 좋은 효과를 보임
- Domain 내 data로 detector를 훈련하는 Detection Approach 방법
- Voting strategy를 통해 여러 reader의 참여를 유도하여 답을 예측하는 방법

5. Conclusion

- 이 연구에서는 misinformation 자동 생성을 위해 LLM을 활용하는 것의 실용성에 대한 평가를 제시했음
- 또한, ODQA system과 같이 지식집약적인 어플리케이션에 대한 영향을 조사했음

- Malicious user가 ODQA system에 일부러 misinformation을 도입하는 scenario를 시뮬레이션함
  - 이를 통해, system이 misinformation에 매우 취약해 성능이 크게 저하된다는 것을 발견함

- 이 연구에서는 이러한 위험에 대응하여 LLM 오용을 완화하기 위한 세 가지 방법을 제안했음

Thank You

감사합니다.