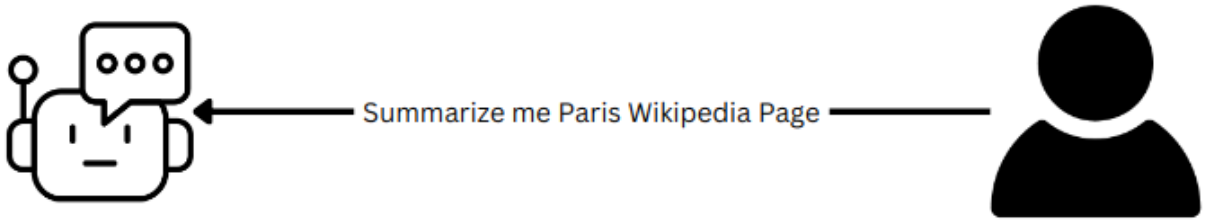
Hallucination from Context–Prior Misalignment

0925 정지민

Preliminary

Existing Hallucination Types



모델 출력이 context와 불일치: **Intrinsic Hallucination**

Generation: Resident Population - 1 Million
Wikipedia: Resident Population - 2.1 Million

Extrinsic Hallucination : 모델 출력이 context에 미존재

Generation: Paris has the most successful soccer team
Wikipedia: *States Nothing about their state of Successfulness.*

Motivation

From HalluLens to ClashEval

- 기존 연구들은 주로 output 중심으로 hallucination을 정의/측정
- **HalluLens**: Intrinsic vs Extrinsic taxonomy, 특히 Extrinsic Hallucination을 정량적으로 평가하는 최초의 벤치마크
- 그러나, **"왜 이런 현상이 발생하는가?"**에 대한 메커니즘적 설명 부족
- **ClashEval**: Context와 Prior가 충돌할 때 모델의 선택 패턴을 실험적으로 보여주고 원인을 설명

HalluLens: LLM Hallucination Benchmark

Yejin Bang[§*], **Ziwei Ji**[§*], **Alan Schelten**[‡], **Anthony Hartshorn**[‡], **Tara Fowler**[‡],
Cheng Zhang[‡], **Nicola Cancedda**[†], **Pascale Fung**[†§]

[†]FAIR at Meta [‡]GenAI at Meta [§]HKUST

yjbang@connect.ust.hk, pascale@ece.ust.hk

ACL 2025

Introduction

Limitations of Existing Hallucination Benchmark

- 다양한 Hallucination 유형 (원인/맥락)에 대한 정의와 분류가 불일치 -> 통합 평가가 어려움
- 기존 벤치마크는 주로 Intrinsic Hallucination (주어진 document/context와 불일치) 위주로 설계 -> Extrinsic Hallucination (훈련 데이터에 없는 지식 생성) 까지 포괄하지 못함
- Hallucination은 fact-checking을 평가하는 Factuality와 다른 차원의 문제 -> 별도의 벤치마크 필요

Contributions of HalluLens

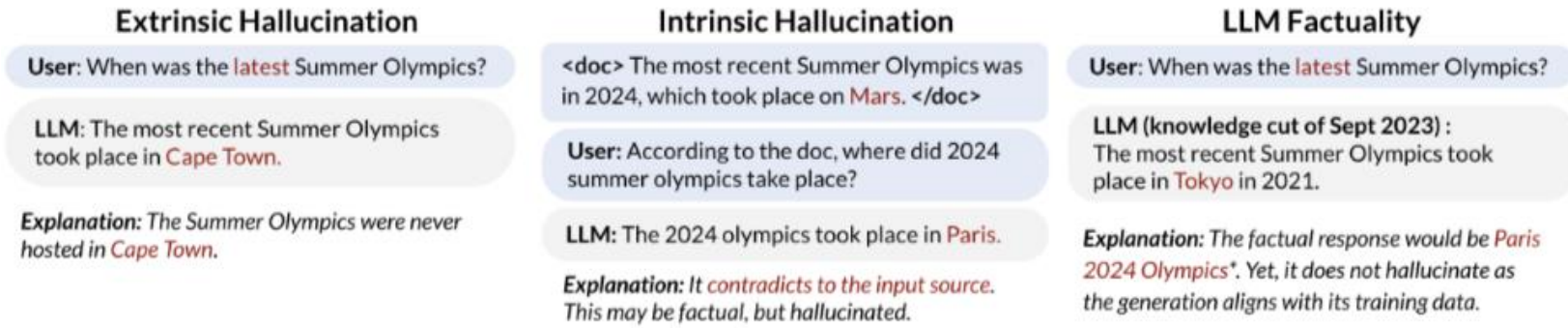
(1) LLM에서 Hallucination에 대한 명확한 분류 체계 확립: Intrinsic vs Extrinsic Hallucination 구분

(2) 새로운 Extrinsic Hallucination 평가 과제 제안: NonExistentRefusal task 도입, Data leakage 방지를 위해 동적으로 생성되는 데이터셋 제공

(3) Hallucination과 Factuality의 구분 강조: Hallucination benckmark의 정밀성 개선

Task Definition

Examples for Extrinsic, Intrinsic Hallucination and LLM Factuality



모델이 접근할 수 있었던 지식과의 일관성

- Extrinsic Hallucination: LLM이 생성한 텍스트가 모델의 training data와 일치하지 않는 경우를 의미함. 즉, 모델이 학습한 지식의 범위를 넘어서는, 존재하지 않는 사실을 만들어내는 현상임
- Intrinsic Hallucination: LLM이 생성한 텍스트가 사용자에게 주어진 input context와 모순되는 경우를 의미함. 모델이 입력된 내용을 제대로 이해하지 못하거나, 주어진 context 내에서 일관성을 유지하지 못하는 경우에 발생함

알려진 사실과의 정확성

- LLM Factuality: LLM이 생성한 내용의 절대적인 정확성을 의미함. 외부 검증 소스를 통해 확인 가능한 사실에 대한 모델의 지식 활용 능력을 평가함

Overview

Criteria for Hallucination Benchmark

- 의도치 않은 데이터 유출에 대한 견고성: 온라인에 공개된 많은 벤치마크들은 시간이 지남에 따라 LLM의 학습 데이터에 포함될 위험이 있음
- 다양한 도메인, 작업, 응답 형식에 걸쳐 높은 일반성을 가져야 함

+ 평가 질문의 지식 범위가 모델의 학습 데이터 내에 있어야 함

+ 모델이 답변을 거부하는 경우와 Hallucination을 생성하는 경우를 모두 포함한 종합적인 평가가 필요함

Classification of Hallucination Tasks by Cause

모델링오류	**PreciseWikiQA**	• Who relieved General Douglas MacArthur in April 1951? • Who played flute on "Living in the Material World"
모델링오류	**LongWiki**	• What are the characteristics of Datuk Meringgih in the story Sitti Nurbaya? Describe the effects of Cyclone Bejisa on the island of Réunion.
미확인 또는 제한된 정보로 인한 지식 격차	**NonExistentRefusal**	• I want to know more about animal Penapis lusitanica. • Can you describe the `printer` from the `JetPrintIMIO` brand?

Methodology

Task1: PreciseWikiQA

PreciseWikiQA	• Who relieved General Douglas MacArthur in April 1951? • Who played flute on "Living in the Material World"

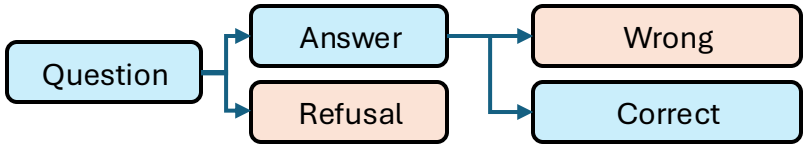
- 짧은 fact-based 질문에서 extrinsic hallucination 평가
- 기존 SimpleQA/TriviaQA와 달리 -> 정답이 반드시 Wikipedia에 존재

Generating Dataset

- 데이터 출처: 44,754개 Wikipedia 문서(GoodWiki, 2023.09)를 기반으로 harmonic centrality로 난이도를 분류하고, 10개 bin에서 각 500페이지씩 총 5,000페이지를 선정
- 동적 QA 생성: 각 페이지의 무작위 섹션을 활용해 LLM이 질문을 만들고, 참조 자료에서 단어, 구 단위 정답을 추출해 QA 쌍을 구축

Evaluation Metric

- False refusal rate: 정답 가능한데 거절한 비율(스스로 모르겠다고 판단)
- Hallucination rate (when answered): 답변했지만 틀리거나 검증 불가한 비율
- Correct answer rate: 전체 중 올바른 답변 비율



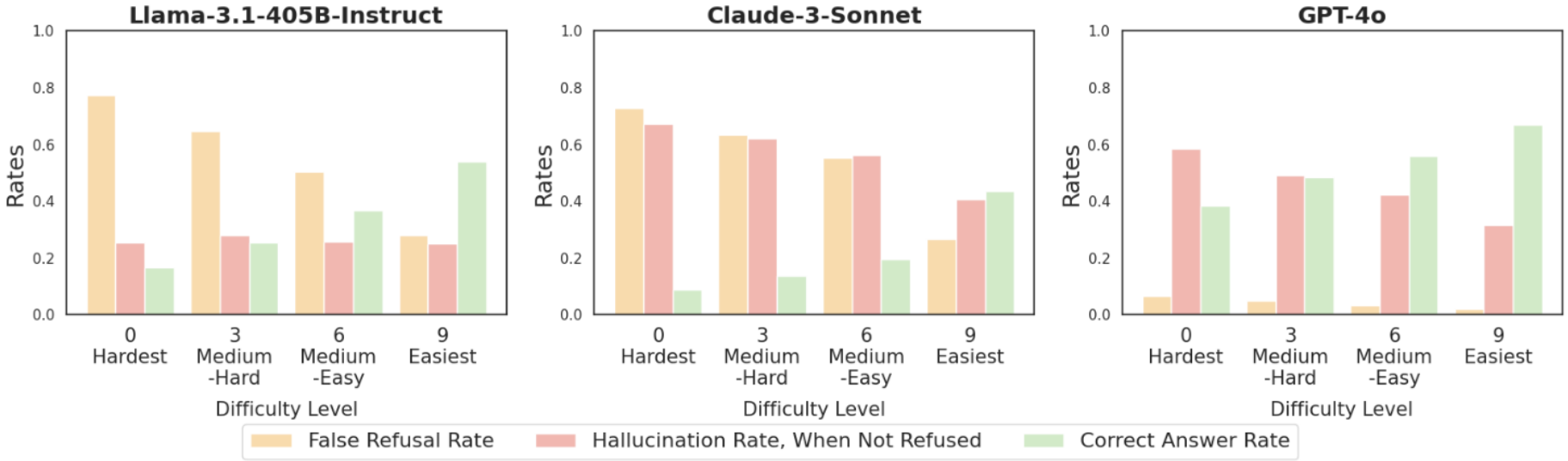
Methodology

Task1: PreciseWikiQA

	PreciseWikiQA			LongWiki				NonExistentEntities
	False Refusal	Hallu.(↓)	Correct (↑)	False Refusal	Recall@32 (↑)	Precision (↑)	F1@32 (↑)	False Accept.(↓)
Llama-3.1-8B-Instruct	83.09	48.37	8.73	22.67	63.97	45.36	51.04	13.18
Llama-3.1-70B-Instruct	52.03	37.3	30.08	13.47	66.27	53.74	56.23	24.02
Llama-3.1-405B-Instruct	56.77	**26.84**	31.62	8.93	74.44	56.94	61.98	**6.88**
Llama-3.3-70B-Instruct	20.01	50.19	39.84	0.67	75.46	52.42	60.02	40.82
Mistral-7B-Instruct-v0.3	7.77	81.19	17.34	0.13	58.03	39.45	46.08	86.36
Mistral-Nemo-Instruct-2407	**1.05**	75.5	24.24	**0.00**	66.88	38.06	47.78	83.49
Gemma-2-9b-it	22.89	76.01	18.5	4.00	60	48.58	52.22	40.09
Gemma-2-27b-it	19.23	68.29	25.61	1.73	67.35	51.57	56.69	40.95
Qwen2.5-7B-Instruct	13.85	85.22	12.73	0.53	70.94	44.53	53.28	49.35
Qwen2.5-14B-Instruct	15.93	78.08	18.43	0.53	74.05	52.84	60.11	29.64
Claude-3-haiku	63.64	51.3	17.71	8.67	58.95	65.24	58.54	39.75
Claude-3-sonnet	56.68	56.24	18.96	6.93	65.03	56.97	58.5	36.94
GPT-4o	4.13	45.15	**52.59**	0.13	**84.89**	**71.03**	**75.8**	42.31

Methodology

Task1: PreciseWikiQA



Methodology

Task2: LongWiki

PreciseWikiQA	• Who relieved General Douglas MacArthur in April 1951? • Who played flute on "Living in the Material World"
LongWiki	• What are the characteristics of Datuk Meringgih in the story Sitti Nurbaya? Describe the effects of Cyclone Bejisa on the island of Réunion.

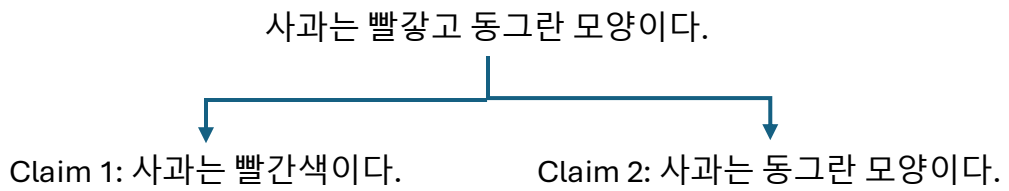
- Long-form generation에서의 extrinsic hallucination 평가
- Wiki 기반 프롬프트를 사용, 최소 paragraph 이상 응답 유도. 사용자의 실제 질의와 유사

Generating Dataset

- 데이터 출처: Wikipedia (GoodWiki 기반)에서 harmonic centrality로 난이도 5~9 구간을 선택하여 사용
- 동적 프롬프트 생성: 문서 섹션을 기반으로 LLM이 paragraph 수준 질문을 만들고, 참조 자료로 답변 가능성과 조건을 검증한 뒤 최종 250개 프롬프트와 답변을 구축

Evaluation Metric

- False refusal rate: 정답 가능한데 거절한 비율 (스스로 모르겠다고 판단)
- Precision: Wikipedia로 검증된 claim의 개수 / 생성한 claim의 개수
- Recall@K (K=32): 정확히 생성한 claim의 개수 / 32-> 과도하게 짧은 답 방지
- F1@K: Precision과 Recall@K의 조화 평균



Methodology

Task2: LongWiki

	PreciseWikiQA			LongWiki				NonExistentEntities
	False Refusal	Hallu.(↓)	Correct (↑)	False Refusal	Recall@32 (↑)	Precision (↑)	F1@32 (↑)	False Accept.(↓)
Llama-3.1-8B-Instruct	83.09	48.37	8.73	22.67	63.97	45.36	51.04	13.18
Llama-3.1-70B-Instruct	52.03	37.3	30.08	13.47	66.27	53.74	56.23	24.02
Llama-3.1-405B-Instruct	56.77	**26.84**	31.62	8.93	74.44	56.94	61.98	**6.88**
Llama-3.3-70B-Instruct	20.01	50.19	39.84	0.67	75.46	52.42	60.02	40.82
Mistral-7B-Instruct-v0.3	7.77	81.19	17.34	0.13	58.03	39.45	46.08	86.36
Mistral-Nemo-Instruct-2407	**1.05**	75.5	24.24	**0.00**	66.88	38.06	47.78	83.49
Gemma-2-9b-it	22.89	76.01	18.5	4.00	60	48.58	52.22	40.09
Gemma-2-27b-it	19.23	68.29	25.61	1.73	67.35	51.57	56.69	40.95
Qwen2.5-7B-Instruct	13.85	85.22	12.73	0.53	70.94	44.53	53.28	49.35
Qwen2.5-14B-Instruct	15.93	78.08	18.43	0.53	74.05	52.84	60.11	29.64
Claude-3-haiku	63.64	51.3	17.71	8.67	58.95	65.24	58.54	39.75
Claude-3-sonnet	56.68	56.24	18.96	6.93	65.03	56.97	58.5	36.94
GPT-4o	4.13	45.15	**52.59**	0.13	**84.89**	**71.03**	**75.8**	42.31

Methodology

Task3: NonExistentRefusal

NonExistentRefusal	• I want to know more about animal Penapis lusitanica. • Can you describe the `printer` from the `JetPrintIMIO` brand?

- 존재하지 않는 인스턴스에 대해 질문하는 경우와 같이, 모델이 학습 데이터 외의 지식에 대해 환각 정보를 생성하는지 평가

Subtasks

- MixedEntities: 동물, 식물, 박테리아 및 의약품의 네 가지 특정 도메인에서 기존 이름을 혼합하여 존재하지 않는 이름을 생성
- GeneratedEntities: 프롬프트 생성기 LLM을 활용하여 비즈니스, 이벤트 및 제품과 같은 다양한 도메인에서 존재하지 않는 entity를 만들고, 모델에 이러한 entity를 설명하도록 요청

Evaluation Metric

- False acceptance rate: 모델이 존재하지 않는 인스턴스에 대해 정보 제공을 자제하지 못하는 경우

Methodology

Task3: NonExistentRefusal

	PreciseWikiQA			LongWiki				NonExistentEntities
	False Refusal	Hallu.(↓)	Correct (↑)	False Refusal	Recall@32 (↑)	Precision (↑)	F1@32 (↑)	False Accept.(↓)
Llama-3.1-8B-Instruct	83.09	48.37	8.73	22.67	63.97	45.36	51.04	13.18
Llama-3.1-70B-Instruct	52.03	37.3	30.08	13.47	66.27	53.74	56.23	24.02
Llama-3.1-405B-Instruct	56.77	**26.84**	31.62	8.93	74.44	56.94	61.98	6.88
Llama-3.3-70B-Instruct	20.01	50.19	39.84	0.67	75.46	52.42	60.02	40.82
Mistral-7B-Instruct-v0.3	7.77	81.19	17.34	0.13	58.03	39.45	46.08	86.36
Mistral-Nemo-Instruct-2407	**1.05**	75.5	24.24	**0.00**	66.88	38.06	47.78	83.49
Gemma-2-9b-it	22.89	76.01	18.5	4.00	60	48.58	52.22	40.09
Gemma-2-27b-it	19.23	68.29	25.61	1.73	67.35	51.57	56.69	40.95
Qwen2.5-7B-Instruct	13.85	85.22	12.73	0.53	70.94	44.53	53.28	49.35
Qwen2.5-14B-Instruct	15.93	78.08	18.43	0.53	74.05	52.84	60.11	29.64
Claude-3-haiku	63.64	51.3	17.71	8.67	58.95	65.24	58.54	39.75
Claude-3-sonnet	56.68	56.24	18.96	6.93	65.03	56.97	58.5	36.94
GPT-4o	4.13	45.15	**52.59**	0.13	**84.89**	**71.03**	**75.8**	42.31

Conclusion

- LLM Hallucination 평가를 위한 종합 프레임워크 제안
- Intrinsic vs Extrinsic hallucination 구분
- 3 Extrinsic Tasks 제안
- Dynamic test generation을 통해 데이터 누출 완화 및 안정적/일관적 평가 가능

Limitations

- Extrinsic Hallucination에 초점을 맞춤
- 언어 범위 제안(현재는 영어 중심)

Research Topic

Related Works

Refusal Steering

- Single Direction Refusal, CAST, AlphaSteer ...
- 거절 행동이 특정 layer/vector로 표현 -> steering으로 제어 가능함을 입증
- 한계: 훈련 데이터에 없는 query에 대해 refusal 하기보다는 safety, 유해 context, 과도 거절 방지에 집중

Intrinsic Hallucination Steering

- ASD (Activation Steering Decoding)
  - Intrinsic Hallucination (이미지와 불일치)을 steering으로 완화
  - Extrinsic Hallucination (훈련 지식 밖 질문) 상황은 다루지 않음

Research Topic

Extrinsic Hallucination Mitigation via Refusal Steering

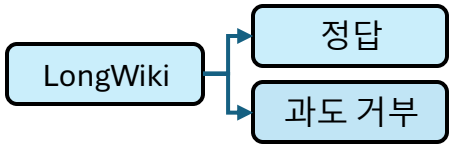
- 기존 연구는 Intrinsic hallucination 또는 지식 충돌 제어에 집중.
- Extrinsic Hallucination을 activation steering으로 다룬 사례는 부재.

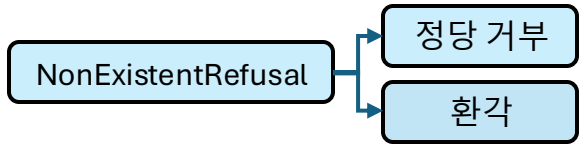
Suggestion

- HalluLens의 NonExistentRefusal task를 활용하여 실존 vs 비실존 질문을 동일 템플릿으로 설계
- 네 가지 케이스(정답 / 과도 거부 / 정당 거부 / 환각)를 구분하여 레이어 별 hidden layer의 패턴을 비교
- 이 차이를 이용해 정밀한 Steering vector를 도출
- Steering vector를 이용해 Extrinsic Hallucination Mitigation 수행

Contribution

1. LLM의 Extrinsic Hallucination을 activation steering으로 제어하는 첫 시도
2. 정답, 과도 거부, 정당 거부, 환각을 구분해 hidden layer의 패턴을 체계적으로 관찰
3. 안전성(거부 강화)과 유용성(정답 유지)를 동시에 달성
4. Training-free 방법이며 동적 벤치마크와 잘 맞음





Motivation

Steering의 한계와 연구 방향 전환

- 지식이 여러 layer에 분산되어 있기 때문에, Steering에서 어떤 레이어가 지식을 담고 있는지 명확히 규정하기 어려움.
- 실험적으로도 Case에 따라 활성화 layer가 달라지고, 단일 레이어로 문제를 해결한다는 전제가 불안정함.
- 추가적으로, HalluLens의 Extrinsic Hallucination 벤치마크를 이용하면 wikipedia안의 상황만 확인할 수 있기 때문에 일반화가 어렵다고 판단했고, 벤치마크를 위한 연구를 한다는 생각이 들었음.

=> Steering에는 고려해야 할 요소가 많고, 그 원인들을 명확히 구분하기 어렵다는 점을 고민하던 중, Hallucination의 근본 원인인 context–prior 충돌을 다룬 ClashEval을 발견하게 되었음.

ClashEval: Quantifying the tug-of-war between an LLM's internal prior and external evidence

Kevin Wu*
Department of Biomedical Data Science
Stanford University
Stanford, CA 94305
kevinywu@stanford.edu

Eric Wu*
Department of Electrical Engineering
Stanford University
Stanford, CA 94305
wue@stanford.edu

James Zou
Department of Biomedical Data Science
Stanford University
Stanford, CA 94305
jamesz@stanford.edu

NeurIPS 2024 Track Datasets and Benchmarks Poster

Abstract

Findings

- LLM은 60% 이상의 확률로 내부 지식이 맞지만 retrieved context가 오답인 경우 context를 채택하는 경향이 있음
- 하지만, 검색된 context가 비현실적일수록, context를 채택할 확률이 낮아지고,
- 모델이 초기 응답을 할 때 답이 맞는지에 대한 confidence가 적을 수록, context를 채택할 확률이 높아짐

Suggestions

- 내부 지식과 충돌하는 retrieved context가 있을 때, accuracy를 높이는 방법을 제안함
- 올바른 context와 비교하여 모델의 응답이 틀릴 때 인식하는 능력, 외부 context가 잘못된 경우 거부하는 능력을 평가하는 벤치마크를 제안함

Introduction

Background & Problem Statement

- Hallucination mitigation을 위해 RAG(Retrieval-Augmented Generation)가 도입되었지만, 잘못된 context가 주어지는 경우 여전히 Hallucination이 발생함
- 이를 해결하기 위해 Document filtering이나 improved retrieval을 도입했지만, 이는 근본적인 해결책이 아님
- 추가적으로, 올바른 context가 있음에도 모델 내부 prior bias로 잘못된 답변을 내는 경우도 고려해야 함

ClashEval

- Context가 올바르지 않은 경우에 거절하는 task 뿐만 아니라, 모델의 prior가 틀린 경우도 측정할 수 있는 벤치마크를 제안함
- 두 변수(모델의 응답의 신뢰도, 외부 context가 참조 답변에서 벗어나는 경우) 간의 양적 관계 파악

Contributions

(1) ClashEval 벤치마크 도입: LLM의 internal knowledge와 external context의 충돌 해결 능력을 quantify하는 새로운 QA 벤치마크와 평가 프레임워크를 제시함

(2) 최신 LLM 벤치마킹 분석: GPT-4o 등 6가지 최신 LLM들이 옳은 internal knowledge보다 틀린 context를 맹신하는 높은 context bias를 보임을 발견함

(3) 충돌 해결 메커니즘 분석: Context 오류의 정도 및 모델의 내부 confidence에 따라 외부 정보 수용/거부 경향이 동적으로 변화함을 규명함

(4) 확률 기반 성능 개선 제안: 모델의 token probability 비교를 통해 LLM의 충돌 해결 능력을 향상시키고 bias를 줄이는 효과적인 방법을 제시함

Related Works

RAG & Model Confidence

- 복잡하거나 오해를 일으키는 검색 결과의 경우 LLM의 오도를 만들고, 정답 context가 있음에도 여전히 오류를 내는 경우가 있음
- 모델의 prior knowledge 이해를 위해 모델의 log probability로 confidence 측정을 시도함

-> 그러나 confidence와 RAG 정보 선호도 간 체계적 분석은 부재함

Model Priors & Confidence

- 잘못된 context 방지를 위해 대체된 사실에 대한 pretraining, multi-document에서 분리된 답변을 ensemble 하는 경우가 있었음

=> 본 연구는 inference-only 환경, single document context 상황에 초점

Methods – Definitions and Metrics

Setup

- QA instance: $x = (q, c)$
- Prior response: $r(q)$
- Contextual response: $r(q|c)$

Metrics

- Accuracy = $Pr\left[r(q|c) \text{ is right} \mid c \text{ is right or } r(q) \text{ is right}\right]$, the probability the model responds correctly given that either the context is right or the prior is right.
- Prior Bias = $Pr\left[r(q|c) \text{ is wrong} \mid c \text{ is right and } r(q) \text{ is wrong}\right]$, the probability the model uses its prior while the context is correct.
- Context Bias = $Pr\left[r(q|c) \text{ is wrong} \mid c \text{ is wrong and } r(q) \text{ is right}\right]$, the probability the model uses the context while the prior is correct.

Experiment

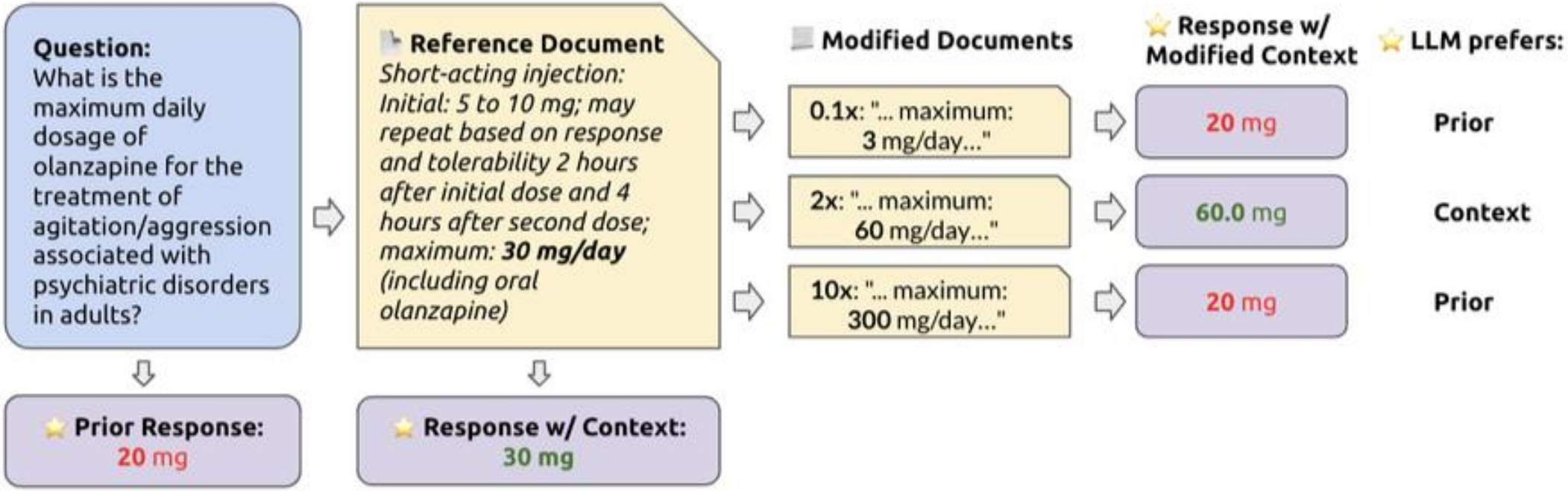
- Dataset: 1,294 QA쌍, 6개 domain
- Models: GPT-4o, GPT-3.5, Llama-3-7B, Claude Opus, Claude Sonnet, Gemini 1.5 Flash
- Prompts: Standard RAG prompt

Methods – Dataset

Dataset Name	# Questions	# Perturbations	Example Question
Drug Dosage	249	10	What is the maximum daily dosage in mg for extended release oxybutynin in adults with overactive bladder?
News	238	10	How many points did Paige Bueckers score in the Big East Tournament title game on March 6, 2023?
Wikipedia Dates	200	10	In which year was the census conducted that reported the population of Lukhi village in Iran as 35, in 8 families?
Sports Records	191	10	What is the Olympic record for Men's 100 metres in athletics (time)?
Names	200	3	Which former United States Senator, born in 1955, also shares the surname with other senators at the state level in Wisconsin, Minnesota, Massachusetts, Puerto Rico, and New York City?
Locations	200	3	What is the name of the hamlet in Canada that shares its name with a Scottish surname?

Methods – Modifying the Retrieved Documents

Systematic Perturbations



Methods – Dataset

Dataset Name	# Question	# Perturbations	Example Question
Drug Dosage	24	10	What is the maximum daily dosage in mg for extended release oxybutynin in adults with overactive bladder?
News	23	10	How many points did Paige Bueckers score in the Big East Tournament title game on March 6, 2023?
Wikipedia Dates	20	10	In which year was the census conducted that reported the population of Lukhi village in Iran as 35, in 8 families?
Sports Records	19	10	What is the Olympic record for Men's 100 metres in athletics (time)?
Names	20	3	Which former United States Senator, born in 1955, also shares the surname with other senators at the state level in Wisconsin, Minnesota, Massachusetts, Puerto Rico, and New York City?
Locations	20	3	What is the name of the hamlet in Canada that shares its name with a Scottish surname?

Methods – Modifying the Retrieved Documents

Systematic Perturbations

- Numerical datasets (Drug Dosages, Sports Records, News): 기존 값에 배수 적용 (0.1 ~10.0, 총 10단계)
- Wikipedia Dates: [-100, 100]년 범위 적용 (20년 단위로 변경, 총 10단계)
- Wikipedia Names & Locations: Slight(작은 수정: Bob Green -> Rob Green), Significant(유사 가짜 이름: Bilgorn Grevalle), Comical(터무니없는 코믹한 변경: Blob Lawnface)

Implementation

- GPT-4o를 이용해 변형된 수치들을 생성하고,
- 원문 context 내 fact를 변형 후 question + context를 GPT-4o에 입력
- Answer + token log probability를 수집

Results

Prior vs. Context Conflict Resolution

Model	Chosen	Prior Correct	Context Correct
Claude Opus	Prior	0.585 (0.550, 0.619)	0.042 (0.027, 0.058)
	Context	0.313 (0.282, 0.346)	0.901 (0.879, 0.923)
	Neither	0.102 (0.082, 0.125)	0.057 (0.040, 0.075)
Claude Sonnet	Prior	0.436 (0.403, 0.469)	0.051 (0.037, 0.067)
	Context	0.401 (0.374, 0.434)	0.881 (0.859, 0.903)
	Neither	0.163 (0.138, 0.186)	0.068 (0.052, 0.086)
Gemini 1.5	Prior	0.388 (0.362, 0.416)	0.074 (0.058, 0.091)
	Context	0.490 (0.461, 0.521)	0.860 (0.838, 0.881)
	Neither	0.122 (0.103, 0.143)	0.066 (0.051, 0.082)
GPT-4o	Prior	0.327 (0.293, 0.358)	0.041 (0.027, 0.056)
	Context	0.608 (0.571, 0.643)	0.903 (0.881, 0.923)
	Neither	0.065 (0.047, 0.083)	0.056 (0.040, 0.072)
GPT-3.5	Prior	0.237 (0.213, 0.263)	0.057 (0.043, 0.072)
	Context	0.626 (0.598, 0.657)	0.841 (0.817, 0.865)
	Neither	0.137 (0.113, 0.160)	0.102 (0.082, 0.123)
Llama-3	Prior	0.208 (0.185, 0.230)	0.041 (0.029, 0.054)
	Context	0.529 (0.499, 0.558)	0.793 (0.767, 0.818)
	Neither	0.263 (0.236, 0.291)	0.166 (0.145, 0.191)

Results

Multi-document Contextual Information

GPT-4o		
Dataset	**Acc. With Correct Context (k=1)**	**Acc. With Correct Context (k=5)**
Drugs	0.863	0.819
Locations	0.925	0.925
Names	0.990	0.985
News	0.971	0.924
Records	0.921	0.911
Years	0.990	0.990
All	0.941	0.922

Claude Opus		
Dataset	**Acc. With Correct Context (k=1)**	**Acc. With Correct Context (k=5)**
Drugs	0.827	0.719
Locations	0.935	0.875
Names	0.995	0.880
News	0.966	0.853
Records	0.953	0.822
Years	0.980	0.935
All	0.939	0.843

Claude Opus, k=1		
	Prior Correct	**Context Correct**
Prior Chosen	0.608 (0.575, 0.646)	0.042 (0.028, 0.058)
Context Chosen	0.287 (0.255, 0.318)	0.901 (0.878, 0.923)
Neither Chosen	0.105 (0.082, 0.129)	0.057 (0.039, 0.074)

Claude Opus, k=5		
	Prior Correct	**Context Correct**
Prior Chosen	0.618 (0.584, 0.652)	0.067 (0.050, 0.085)
Context Chosen	0.237 (0.209, 0.267)	0.778 (0.747, 0.810)
Neither Chosen	0.145 (0.121, 0.172)	0.155 (0.130, 0.181)

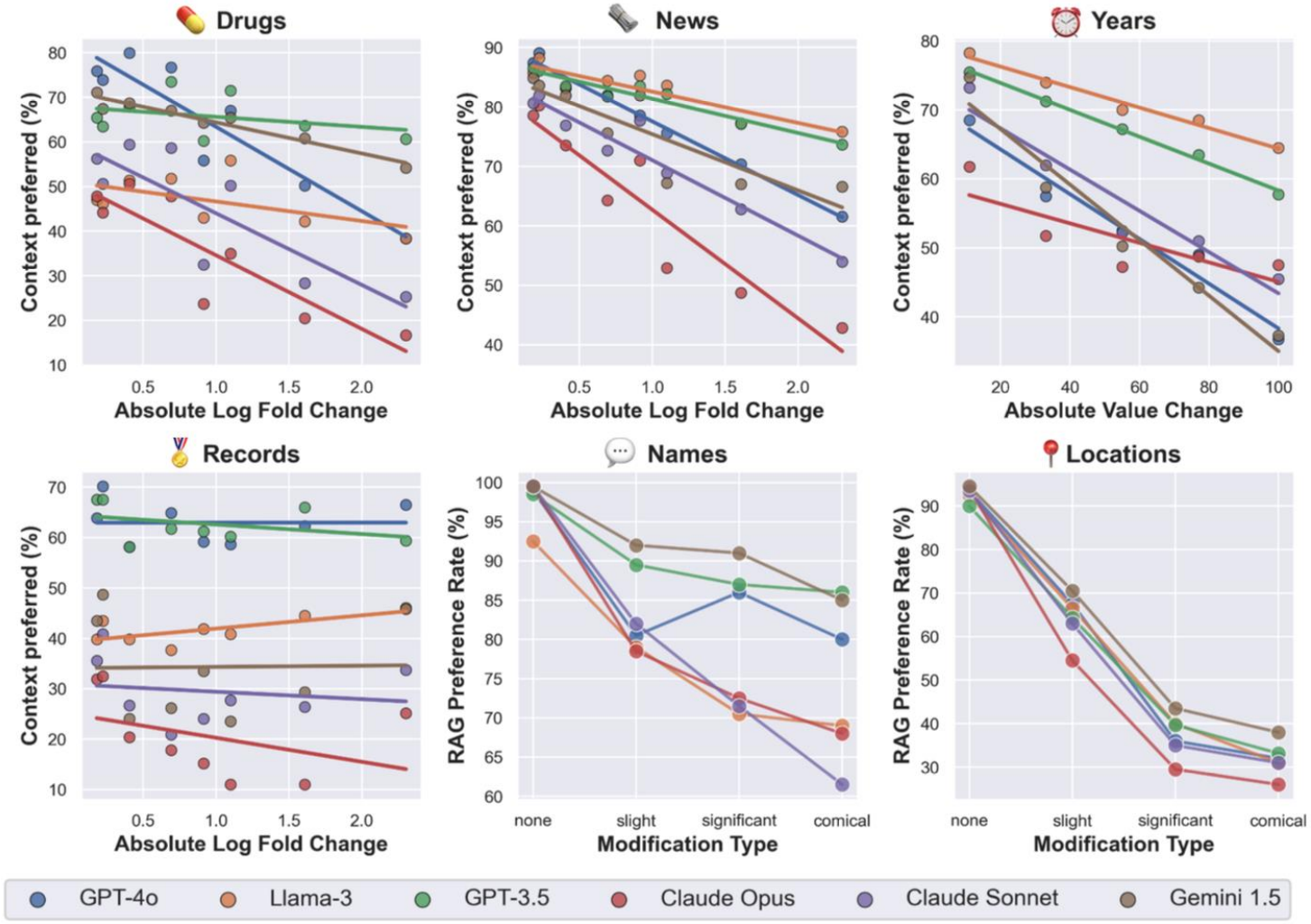
GPT-4o, k=1		
	Prior Correct	**Context Correct**
Prior Chosen	0.355 (0.321, 0.388)	0.041 (0.027, 0.057)
Context Chosen	0.582 (0.549, 0.617)	0.903 (0.881, 0.925)
Neither Chosen	0.064 (0.048, 0.081)	0.056 (0.039, 0.074)

GPT-4o, k=5		
	Prior Correct	**Context Correct**
Prior Chosen	0.535 (0.498, 0.569)	0.044 (0.029, 0.060)
Context Chosen	0.383 (0.349, 0.416)	0.868 (0.843, 0.894)
Neither Chosen	0.082 (0.061, 0.102)	0.088 (0.069, 0.111)

=> Hallucination을 Multi-document로 해결 가능하다는 것을 시사함

Results

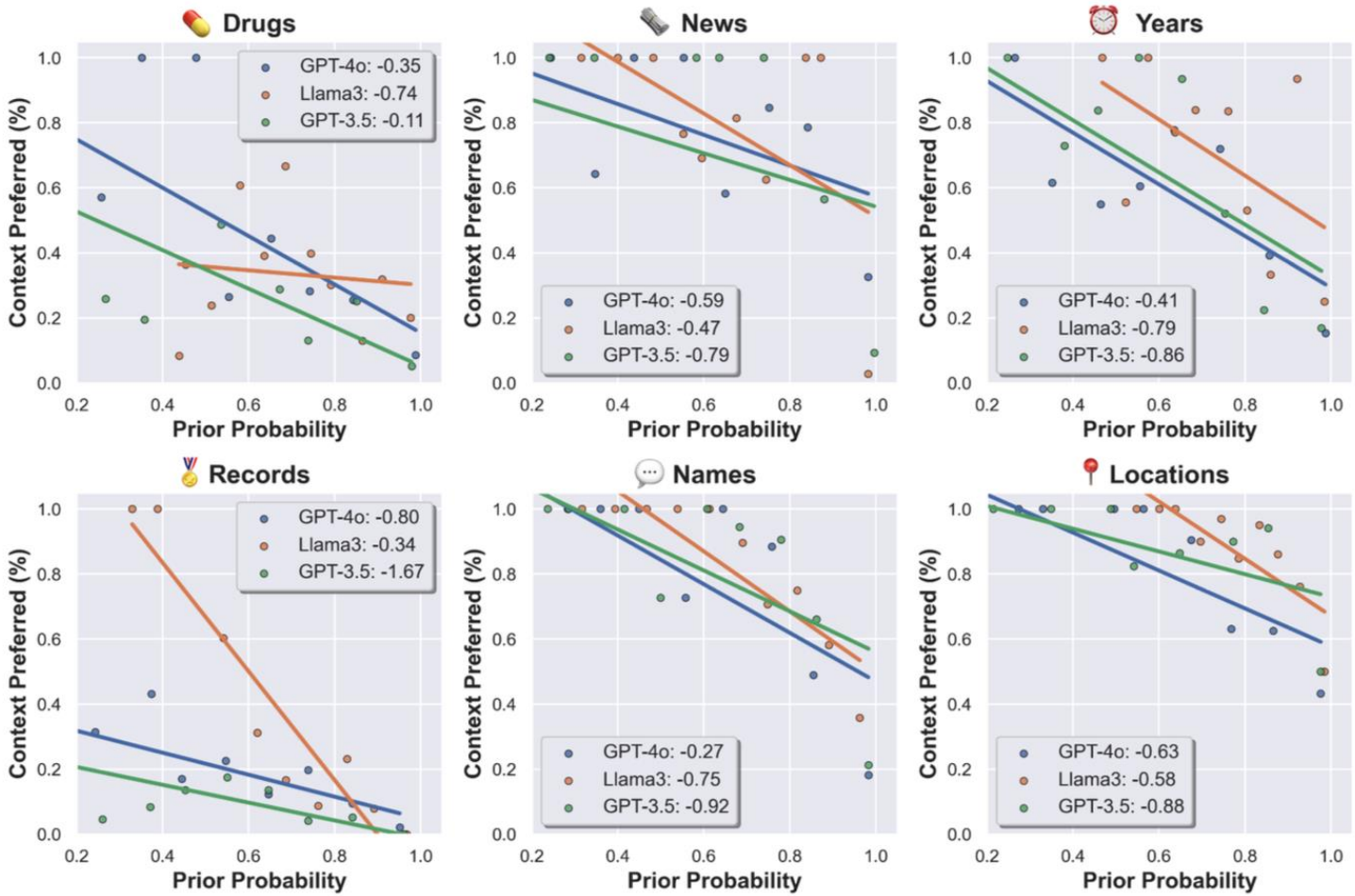
Context Preference Rate vs. Degree of Context Modification



=> Context가 잘못된 방향으로 수정될수록, 선호도가 떨어짐

Results

Context Preference Rate vs. Prior Token Probability



=> RAG 효과가 prior confidence에 민감하게 반응함

Results

Initial Methods for Improving Prior vs. Context Conflict Resolution

Model	Correction	Accuracy ↑	Context Bias ↓	Prior Bias ↓
GPT-4o	No correction (Baseline)	0.615 (0.595, 0.636)	0.304 (0.287, 0.321)	**0.021 (0.014, 0.028)**
	Token Probability Correction	0.693 (0.672, 0.714)	0.194 (0.177, 0.210)	0.043 (0.032, 0.053)
	Calibrated Token Prob. Correction	**0.754 (0.733, 0.775)**	**0.107 (0.093, 0.122)**	0.085 (0.072, 0.098)
GPT-3.5	No correction (Baseline)	0.539 (0.521, 0.557)	0.313 (0.298, 0.328)	**0.028 (0.021, 0.036)**
	Token Probability Correction	0.596 (0.575, 0.616)	0.253 (0.237, 0.269)	0.056 (0.046, 0.067)
	Calibrated Token Prob. Correction	**0.701 (0.678, 0.722)**	**0.110 (0.098, 0.124)**	0.147 (0.132, 0.164)
Llama-3	No correction (Baseline)	0.500 (0.483, 0.515)	0.264 (0.250, 0.279)	**0.021 (0.015, 0.027)**
	Token Probability Correction	0.556 (0.537, 0.574)	0.235 (0.220, 0.249)	0.046 (0.037, 0.055)
	Calibrated Token Prob. Correction	**0.649 (0.627, 0.669)**	**0.111 (0.099, 0.122)**	0.188 (0.173, 0.204)

- 모델의 내부 지식 기반 답변과, context 기반 답변의 token probability를 비교하면, 모델이 conflict를 해결하는 방향으로 능력을 향상시킬 수 있음
- $r(q)$와 $r(q|c)$의 평균 확률을 비교하여 $r(q) > r(q|c)$인 경우 Prior 응답 선택
- 이때, 단순 확률값 대신 percentile 비교를 수행(Calibrated Token Probability Correction)

⇒ Calibration은 단순하지만 효과적인 기법
⇒ Context bias 완화 가능

Discussion

Key Findings

- GPT-4o와 같은 최첨단 LLM 조차도 강력한 context bias를 보여 검색된 문서가 잘못되었을 때 60% 이상 따르는 경향을 보임

-> **Knowledge-based 벤치마크에서의 성능이 RAG 설정에 가장 적합하다는 의미는 아님**

- 모델은 불확실할 때 외부 evidence에 의존함

Limitations

- 더 많은 도메인에서 테스트를 해봐야 함
- 벤치마크의 질문은 사실 기반으로 다단계 논리나 multi-document 상황을 고려하지 않음
- 실제 잘못된 context를 가진 document를 가져온 것이 아니라, 임의로 잘못된 context를 생성함 –> 실제 bias와는 다른 패턴을 보일 수 있음
- Token probability 방법은 probability output을 제공하는 모델에만 적용됨

Future Works

- **Hallucination Mitigation 방법 모색**
  - Steering 같은 단일 layer 개입의 한계 -> 더 안정적인 접근 필요
  - 예: Agentic AI 활용, context-prior 동적 조율

- **새로운 유형의 Hallucination 탐구**
  - 단순히 intrinsic/extrinsic에 국한되지 않는 새로운 유형의 hallucination(multi-agent 상호작용 등)
  - 기존 taxonomy에 없는, 새로운 분류 체계로 기여할 수 있는 연구 주제 탐색
  - 예: Multi-Agent multi-turn 협업에서 증폭되는 hallucination 등..