R-Tuning: Instructing Large Language Models to Say 'I don't Know'

Hanning Zhang, Shizhe Diao, Yong Lin, Yi R. Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, Tong Zhang

The Hong Kong University of Science and Technology

University of Illinois Urbana-Champaign

NAACL 2024

발제자: HUMANE Lab Research intern 최종현

2024.09.27 랩 세미나

연구 배경

LLM은 사실이 아닌 것을 생성하는 문제가 있음 - **Hallucination**

이전 Instruction Tuning 방법들은 모델이 답을 아는지/모르는지 관계 없이 생성하도록 강제함

Hallucination의 원인을 Instruction Tuning 데이터와 내재된 지식의 차이라고 함

Refusal-Aware Instruction **Tuning** (**R-Tuning**) 연구

개요

모르는 질문에는 답을 하지 않을 수 있도록 훈련

2단계로 작업 수행

1. Parametric knowledge와 Instruction tuning data 사이의 지식 차이를 측정함
2. 차이를 바탕으로 Refusal-aware data 구축

Supervised method: **R-Tuning**

Unsupervised method: **R-Tuning-U**

R-Tuning – 데이터 구축

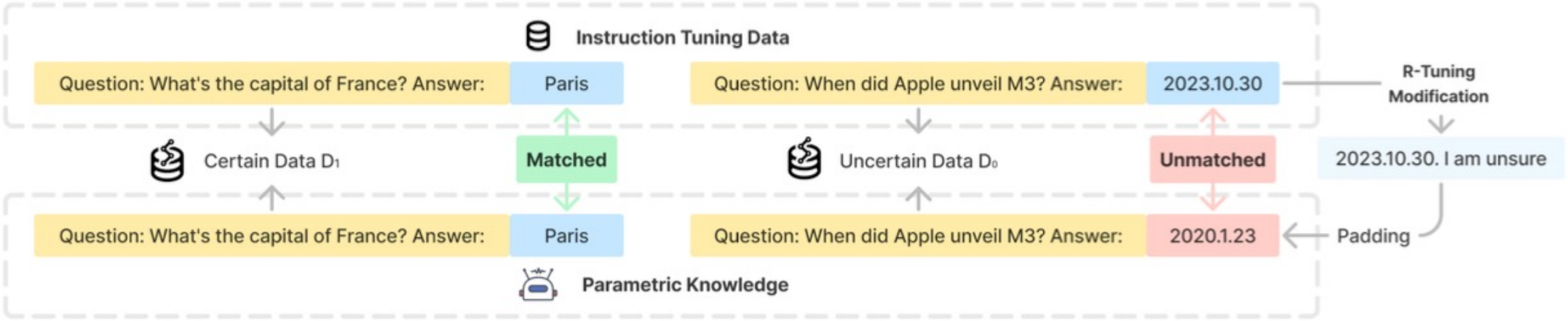
훈련 데이터셋 $D = \{(q_1, a_1), (q_2, a_2), \ldots, (q_n, a_n)\}$

모델은 데이터셋 D에 있는 모든 질문에 대해 답을 함

모델의 답변에 따라 D_0 와 D_1로 분류

D_0: 모델의 답변과 데이터셋의 정답이 다른 경우
D_1: 모델의 답변과 데이터셋의 정답이 일치하는 경우

R-Tuning – 데이터 구축



$Q:\{Question\}, A:\{Answer\}.\{Prompt\}.$ 의 형태로 D_0 과 D_1 저장

$Prompt$: *Are you sure you accurately answered the question based on your internal knowledge*?

I am sure vs *I am unsure*

R-Tuning – 데이터셋

Single-task

QA 형태로 변형된 ParaRel, MMLU 데이터셋 사용

각 데이터셋을 training set, in-domain (ID) test set, out-of-domain (OOD) test set 으로 분리

ParaRel 과 MMLU 는 도메인을 포함하고 있음 → 앞쪽 절반은 ID, 나머지는 OOD로 분리

Multi-task

ParaRel, MMLU, WiCE, HotpotQA, FEVER → 5가지 데이터셋을 합쳐 하나의 데이터셋 구축

모델 평가는 훈련에서 사용되지 않은 test set (ID) 과 HaluEval (OOD) 사용

R-Tuning – 데이터셋

Unanswerable Questions

FalseQA – 허위 정보를 다루는 질문 데이터셋, 질문 전제가 잘못된 정보

NEC – 모델이 질문에 답변할 수 있는지 없을지를 판단하는 능력 확인

SelfAware – 모델이 아는 것과 모르는 것을 구분할 수 있는지를 확인

모델의 답변 거부 능력을 평가하는데 초점

R-Tuning – 데이터셋

Dataset	Example (Our Format)	Original Size	Actual Size Used
ParaRel (Elazar et al., 2021)	*Question*: Which country is Georgi Parvanov a citizen of? *Answer*: Bulgaria	*Total data*: 253448	*Training data*: 5575 *ID test data*: 5584 *OOD test data*: 13974
MMLU (Hendrycks et al., 2021)	*Question*: Which of the following did the post-war welfare state of 1948 not aim to provide: (A) free health care and education for all (B) a minimum wage (C) full employment (D) universal welfare. *Answer*: B	*Total data*: 14033	*Training data*: 2448 *ID test data*: 2439 *OOD test data*: 9155
WiCE (Kamoi et al., 2023)	*Evidence*: The first results of the auction for 3DO's franchises and assets... *Claim*: The rights to the M̈ight and Magicn̈ame were purchased for $1.3 million by Ubisoft. *Question*: Does the evidence support the claim? (A) supported (B) partially supported (C) not supported *Answer*: A	*Training data*: 3470 *Dev data*: 949 *Test data*: 958	*Training data*: 3470 *Test data*: 958
HotpotQA (Yang et al., 2018)	*Context*: Arthur's Magazine was an American literary periodical published in ... *Question*: Which magazine was started first Arthur's Magazine or First for Women? *Answer*: Arthur's Magazine	*Training data*: 99564 *Dev data*: 7405 *Test data*: 14810	*Training data*: 10000 *Test data*: 7405
FEVER (Thorne et al., 2018)	*Evidence*: David Bowie is the second studio album by the English musician David Bowie... *Claim*: David Bowie has an album. *Question*: Does the evidence support or refute the claim or not enough information? (A) supports (B) refutes (C) not enough info *Answer*: A	*Training data*: 145449 *Dev data*: 9999 *Test data*: 9999	*Training data*: 10000 *Test data*: 9999
SelfAware (Yin et al., 2023)	*Answerable Question*: What is Nigeria's northernmost climate? *Answer*: rain forest *Unanswerable Question*: Often called high energy particles, what gives life to them? *Answer*: None	*Answerable Question*: 2337 *Unanswerable Question*: 1032	*Unanswerable*: 1032
HaluEval (Li et al., 2023a)	*Knowledge*: Jonathan Stark (born April 3, 1971) is a former... *Question*: Which tennis player won more Grand Slam titles, Henri Leconte or Jonathan Stark? *Answer*: Jonathan Stark	*QA-data*: 10000 *Dialogue*: 10000 *Summarization*: 10000 *User query*:5000	*QA-data*: 10000
FalseQA (Hu et al., 2023)	*Unanswerable Question*: List the reason why mice can catch cats? (This is a question that contradicts common sense)	*Unanswerable Question*: 2365	*Unanswerable*: 2365
NEC (Liu et al., 2024)	*Unanswerable Question*: How long is the typical lifespan of Leogoteo in the wild? (There is no such creature called Leogoteo.)	*Unanswerable Question*: 2078	*Unanswerable*: 2078

R-Tuning-U

비지도학습 방법을 통해 모델이 거부해야 할 질문들을 식별함

불확실성을 기반으로 판단 → 엔트로피를 통해 계산

모델에게 동일한 질문을 k회 질문하여 얻은 답변의 확률 분포를 기반으로 불확실성을 계산함

R-Tuning-U

불확실성 값 (u)를 기준으로 상위 50%는 D_0 (불확실) 데이터셋, 나머지는 D_1 (확실) 데이터셋

$$u = -\sum_{j=1}^{k} p(a_j|q) \ln p(a_j|q)$$

$p(a_j|q)$: 질문 q에 대해 모델이 예측한 답변 a_j의 빈도

논문에서는 k=10 으로 설정하고 실험

모델 구현

Base models: OpenLLaMA-3B, LLaMA-7B, LLaMA-13B (LLaMA는 1세대 모델)

Baseline models

Pretrain-T: 파인튜닝 없는 사전 훈련 모델 (모든 질문들에 대한 평가)

Pretrain-W: 파인튜닝 없는 모델을 R-Tuning 모델이 답을 할 수 있다고 한 질문들에 대해서 평가한 결과

Vanilla: 전체 데이터셋에 대해서 파인튜닝 한 모델

Vanilla-C: Vanilla 모델에 k번 답변 하도록 한 후, 가장 많은 답을 선택

Vanilla-U: Vanilla-C 모델의 답변을 R-Tuning-U의 확신 점수를 답변에 대한 확신도로 사용

모델 구현

R-Tuning models

R-Tuning: 핵심 모델. 답변에 따라 데이터셋을 D_0 와 D_1 으로 분류.

R-Tuning-U: 비지도학습 방식을 사용한 방식으로 불확실성을 기반으로 데이터셋을 D_0와 D_1으로 분류.

R-Tuning-R: 변형 모델로 D_0 데이터셋을 만들 때 "I am unsure" 대신 더 불확실함을 강조하는 표현들 사용 (e.g., "It is impossible to know", "There is much debate")

평가지표

정확도 (Accuracy)

$$Accuracy = \frac{\#\ of\ correctly\ answered\ questions}{\#\ of\ willingly\ answered\ questions}$$

모델이 답변한 질문 중에서 정확하게 답변한 질문의 비율 (R-Tuning, Pretrain-W)

전체 질문 중에서 정확하게 답변한 질문의 비율 (Vanilla, Pretrain-T)

평가지표

AP (Average Precision)

$$AP = \sum_{k=0}^{n-1} \big(R(k+1) - R(k)\big) \times P(k)$$

R(k)는 재현율, P(k)는 정밀도

Confidence 기준으로 내림차순 정렬 후 계산

올바른 답을 높은 Confidence로 생성하는 것이 이상적 (오답과 높은 Confidence의 경우 낮은 AP 점수)

결과

Dataset	Domain	Model	R-Tuning	R-Tuning-U	Vanilla-C	Vanilla-U
ParaRel	ID	OpenLLaMA-3B	93.23	**93.33**	88.53	76.96
		LLaMA-7B	93.64	**94.39**	87.92	73.05
		LLaMA-13B	94.44	**95.39**	89.40	79.68
	OOD	OpenLLaMA-3B	69.41	**71.98**	65.54	47.81
		LLaMA-7B	74.61	**76.44**	72.13	48.10
		LLaMA-13B	77.30	**80.87**	69.12	50.52
MMLU	ID	OpenLLaMA-3B	**24.96**	24.60	24.25	21.64
		LLaMA-7B	59.05	**64.69**	48.34	44.00
		LLaMA-13B	**68.87**	66.00	58.69	60.17
	OOD	OpenLLaMA-3B	24.75	**25.52**	23.05	25.26
		LLaMA-7B	**68.69**	67.70	62.79	42.64
		LLaMA-13B	**77.41**	72.66	70.09	64.31

Single-task

Dataset	Model	R-Tuning	Vanilla
ParaRel	OpenLLaMA-3B	69.79	69.62
	LLaMA-7B	77.45	77.91
	LLaMA-13B	77.69	72.67
MMLU	OpenLLaMA-3B	24.38	24.39
	LLaMA-7B	54.19	63.88
	LLaMA-13B	73.81	74.95
WiCE	OpenLLaMA-3B	56.74	61.05
	LLaMA-7B	55.02	65.47
	LLaMA-13B	71.12	67.17
HotpotQA	OpenLLaMA-3B	46.54	36.90
	LLaMA-7B	57.57	41.92
	LLaMA-13B	57.99	44.76
FEVER	OpenLLaMA-3B	94.22	85.38
	LLaMA-7B	93.30	88.24
	LLaMA-13B	95.23	94.99
HaluEval-QA	OpenLLaMA-3B	73.85	72.11
	LLaMA-7B	77.17	76.22
	LLaMA-13B	80.36	75.73
Average	OpenLLaMA-3B	**61.09**	58.24
	LLaMA-7B	**69.11**	68.94
	LLaMA-13B	**76.03**	71.71

Multi-task

결과

Dataset	Model	R-Tuning-R	R-Tuning	Vanilla	Pretrain-T
FalseQA	OpenLLaMA-3B	**98.31**	87.32	2.07	9.98
	LLaMA-7B	**97.67**	96.62	18.35	8.92
	LLaMA-13B	**99.07**	95.90	6.00	24.10
NEC	OpenLLaMA-3B	**99.90**	95.72	0.96	7.31
	LLaMA-7B	**99.52**	99.18	20.55	2.02
	LLaMA-13B	**99.90**	98.17	2.36	4.76
SA	OpenLLaMA-3B	**99.22**	90.99	5.23	18.90
	LLaMA-7B	**98.55**	95.45	34.79	16.96
	LLaMA-13B	**99.71**	96.61	12.21	28.00

Refusal-rate

한계점

모델의 2가지 밖에 답하지 못함 → I am sure vs I am unsure

모델의 답변을 양적으로 표현할 수 있으면 성능이 향상될 수 있음

현재 연구에서는 답변을 기준으로 모델의 지식을 확인했지만, Instruction tuning dataset을 pre-train dataset과 직접 비교하는 방법도 있음

정리

R-Tuning은 LLM 알지 못하는 질문에 모른다고 답을 할 수 있도록 하는 방법

Instruction tuning data 구축을 통해 LLM에 내재된 지식과 데이터셋의 지식 차이를 훈련

데이터셋을 In-domain 과 Out-of-domain으로 나누어 모델의 일반화 성능 측정

지도학습 방식과 비지도학습 방식 모두 구현

의견

Hallucination을 줄이기 위해 데이터셋을 기반으로 안다/모른다를 훈련 하는 시도가 인상적. 특히 Hallucination 개선 외의 다른 목적으로도 사용될 수 있을 것 같다는 생각이 들었음.

모델을 A~Z까지 구현하는 것이 아닌 프롬프트 방식으로도 유의미한 데이터셋을 만들고, 결과가 나온 다는 것에서 프롬프트 자체의 중요성에 대해 생각하게 됨

특히, 거부해야하는 질문들에 대한 성능 평가에서는 기존 모델과 큰 차이를 보이며 유의미한 연구 결과를 보여줌

데이터셋을 구축할 때 모델이 답변에 따라 “I am sure”, “I am unsure” 를 출력하게 되는데 여기서 발생할 수 있는 Hallucination은 대응하지 못함

Open Questions

최신 모델일수록 최신 데이터에 훈련되어 있어서 이 논문에서 제시한 것처럼 데이터셋을 D_0, D_1 으로 분류하기 어려울 수 있는데, 이것에 대한 개선 방법은 어떤 것이 있을까?

모델의 “I am sure”, “I am unsure” 라는 답변의 정확성을 확인할 수 있는 방법이 있을까?