Can AI Assistants Know What They Don't Know?

Qinyuan Cheng[12*], Tianxiang Sun[12*], Xiangyang Liu[12],
Wenwei Zhang[2], Zhangyue Yin[1], Shimin Li[1], Linyang Li[12], Zhengfu He[1], Kai Chen[2t], Xipeng Qiu[2t]

Fudan University[1], Shanghai AI Laboratory[2]

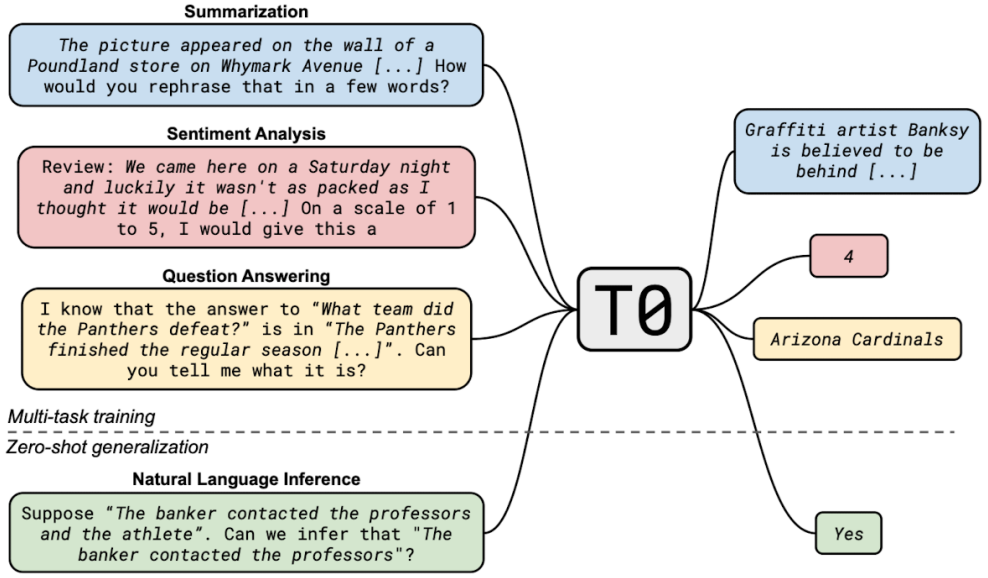
ICML 2024 Poster

2024.10.11

발제자: 윤예준
yeayen789@gmail.com

Background

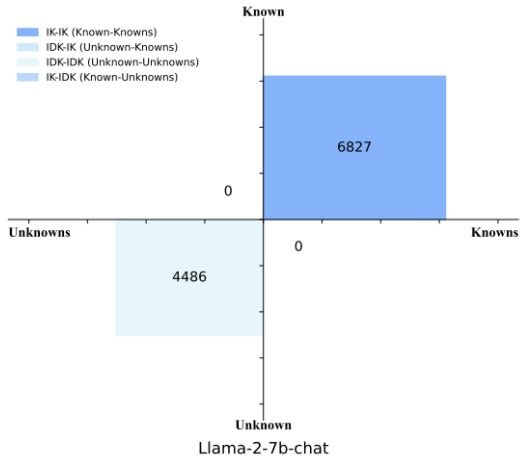
- LLM을 기반으로 하는 AI assistants들은 다양한 task에서 놀라운 성과를 보임
- 그러나 여전히 LLMs은 knowledge-intensive task에서 factual errors를 범함
  → AI assistants의 untruthful responses는 practical applications에서 상당한 위험을 초래할 수 있음



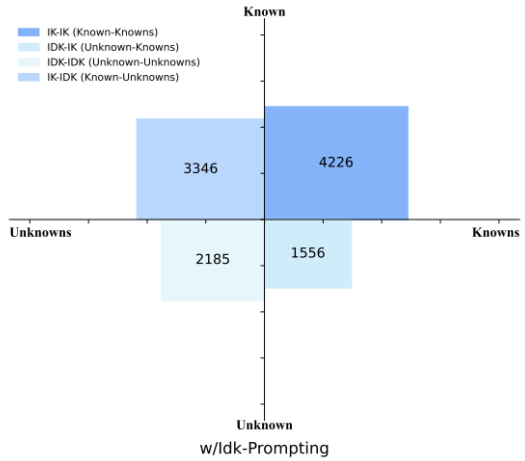
Can AI assistants know what they don't know and express this awareness through natural language?

Knowledge quadrants

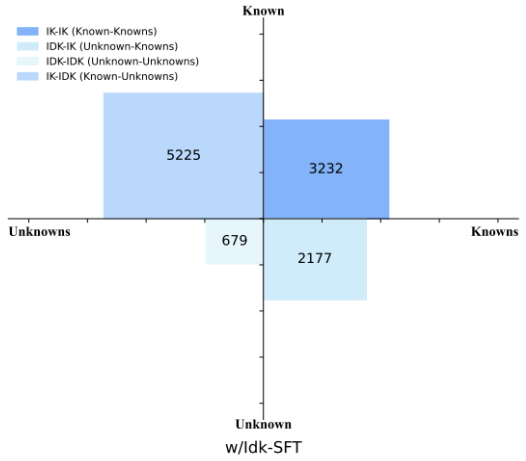
	Unknowns	Knowns
Known	Known Unknowns: Things the AI knows it doesn't know.	Known Knowns: Things the AI knows it knows.
Unknown	Unknown Unknowns: Things the AI doesn't know it doesn't know.	Unknown Knowns: Things the AI doesn't know it knows.



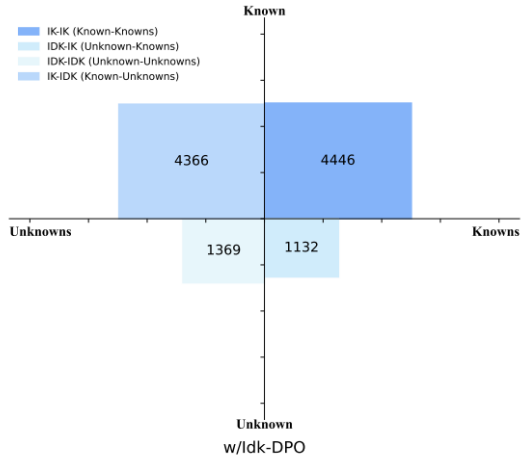
Llama-2-7b-chat



w/Idk-Prompting



w/Idk-SFT



w/Idk-DPO

- "Knowns" represents what the AI actually knows
- "Unknowns" represents what the AI does not actually know
- "Known" represents what the AI believes it knows
- "Unknown" represents what the AI believes it does not know

Findings summarization

- I don’t know (Idk) 데이터셋을 구축 및 Idk를 사용하여 AI assistant가 자신이 알고 있는 것과 모르는 것을 파악하고 모르는 질문은 거절할 수 있도록 함

- Supervised fine-tuning은 모델을 지나치게 조심스럽게 만들어 알려진 질문을 잘못 거부하는 경향 존재
  → Preference-aware optimization은 이를 방지하여 아는 것과 모르는 것을 정확하게 답변하는 비율을 높임

- Idk 데이터셋 구축시 사용하는 Ik threshold는 AI assistant의 동작에 영향을 미침
  → ‘I don’t know’로 표시된 질문이 많을수록 assistant가 질문에 대한 답변을 거부할 가능성이 높아짐

- 일반적으로 Ik threshold이 높을수록 IK-IK, IK-IDK 수가 많아져서 결과적으로 more truthful assistant 됨

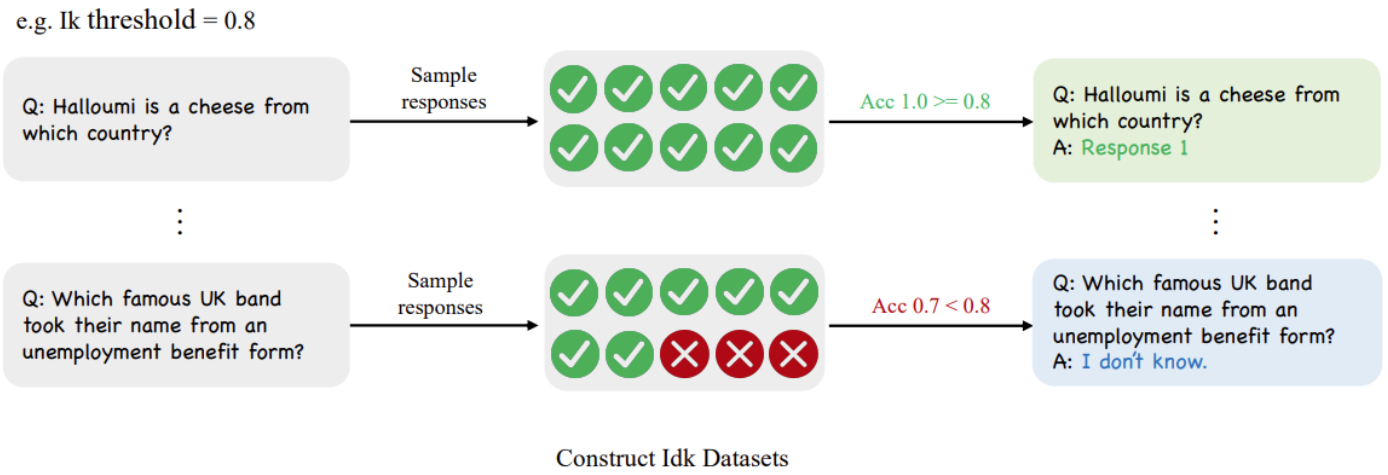
- Larger model은 알고 있는 질문과 모르는 질문을 더 능숙하게 구분할 수 있음

Methodology

Construction of the Idk Dataset

- 평가 방법: lexical matching
- TriviaQA 사용하여 question당 10개 응답 샘플링
- 10개 모두 정답인 경우에만 질문에 대한 답을 안다고 간주
- Refusal to answer template is:

```
This question is beyond the scope of my
knowledge, and I am not sure what the
answer is.
```



Construct Idk Datasets

Methodology

Idk prompting: 입력 질문 앞에 프롬프트를 추가하여 알 수 없는 질문에 모르겠다고 말하도록 직접 지시

```
Answer the following question, and if you
don't know the answer, only reply with "I
don't know": <Question>
```

Training method

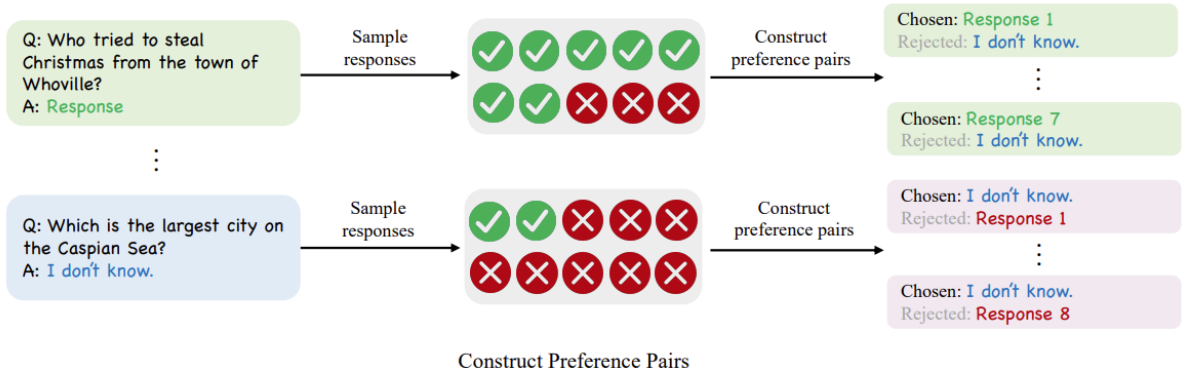
- Idk supervised fine-tuning

$$\mathcal{L}_{SFT} = -E_{(x,y)\sim D}[\frac{1}{N}\sum_{t}^{N}\log p(y_t|x, y_{<t}; \theta)]$$

- Preference-aware optimization
  - Direct preference optimization (DPO)
    - Idk dataset 절반으로 SFT model 학습
    - 나머지 절반 dataset에서 SFT model의 response 수집 (preference pairs 구성 목표)
      - 모델이 정답을 아는 샘플은, 모델이 생성한 정답을 chosen으로 사용하고 I don't know를 rejected로 사용
      - 모델이 정답을 모르는 샘플은, I don't know를 chosen으로 사용하고 모델이 생성한 틀린 응답을 rejected로 사용

$$\mathcal{L}_{DPO} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta\log\frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right)\right]$$

$$\mathcal{L}_{DPO-SFT} = \mathcal{L}_{DPO} + \alpha * \mathcal{L}_{SFT}$$



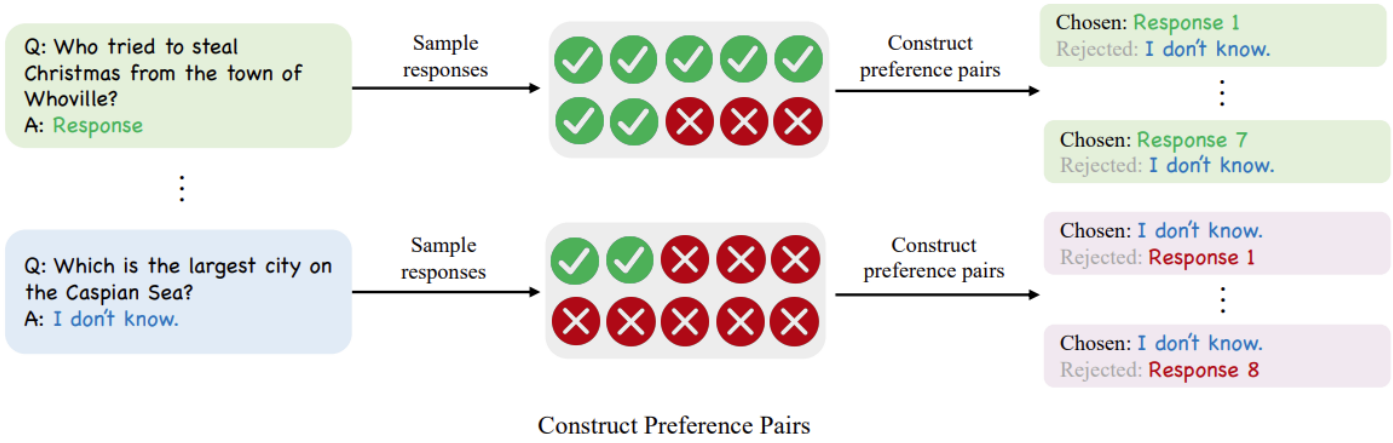
Construct Preference Pairs

Methodology

Training method

- Preference-aware optimization
  - Best-of-n sampling (BoN)
    - Idk dataset 절반으로 SFT model 학습 후 reward model로 초기화
    - 나머지 절반으로 preference pairs 생성 (DPO에서 설명한 방법과 동일)
    - reward model을 pairwise loss로 preference piars로 학습
    - 추론 시 SFT model로 Best-of-10 전략으로 10개의 response 샘플링 후 reward-model로 score 측정
    - 가장 높은 score를 받은 응답을 최정 응답으로 선택

$$\mathcal{L}_{RM} = -E_{(x,y_w,y_l)\sim D}\left[\log \sigma\left(r(x_i, y_w) - r(x_i, y_l)\right)\right]$$



Construct Preference Pairs

Methodology

Training method

- Preference-aware optimization
  - Proximal policy optimization (PPO)
    - BoN 학습 방법에서 사용한 SFT model과 reward 모델을 사용하여 학습

  - Hindsight instruction relabeling (HIR)
    - 아래 instruction을 추가하여 Idk dataset을 다시 레이블링

```
Your current knowledge expression
confidence level is <X>, please answer the
user's question: <Question>
```

    - <X> is the value of model's knowledge expression confidence level ranging from 0 to 1.0

$$Knowledge\_expression\_confidence\_level = 1.1 - Ik\_threshold$$

    - confidence level이 낮을 수록 질문에 대답하는 것을 거부할 가능성이 높음
    - 기존 Idk 데이터셋과 결합하여 supervised fine-tuning 진행
    - instruction relabeling의 이점은 모델을 다시 학습할 필요 없이 conservative or aggressive response strategy 채택할 수 있음

Experiments

- Datasets
  - TriviaQA: ODQA dataset으로 wikipedia와 웹에서 수집한 question-answer pairs로 구성됨
    - Training set: Trivia QA train set의 90% - 78,899 건
    - Validation set: Trivia QA train set의 10% - 8,763건
  - Test set
    - Trivia QA development set 전체 – 11,313
    - Out-of-distribution
      - Natural Questions: real queries from the Google search engine
        - Devset 3,610건 사용
      - ALCUNA: a benchmark to assess LLMs' abilities in new knowledge understanding
        - 기존 entity를 변경하여 new artificial entities를 생성
        - ALCUNA 질문 중 일부를 사용하여 8,857 건 생성

Results

- Metrics
  - IK-IK(Known-Knowns) Rate: 모델이 올바르게 답한 질문의 비율
  - IK-IDK(Know-Unknowns) Rate: 모델이 올바르게 대답을 거부하는 질문 비율
  - TRUTHFUL Rate: IK-IK rate과 IK-IDK rate의 합. 진실한 답을 제공하는 질문의 비율

Table 1. Overall results on the test set of the Idk dataset constructed based on TriviaQA and out-of-distribution test sets.

	TriviaQA			**Natural Questions**			**ALCUNA**
	IK-IK	IK-IDK	TRUTHFUL	IK-IK	IK-IDK	TRUTHFUL	IK-IDK
Idk-Dataset$_{test}$	45.05	54.95	100.00	24.65	75.35	100.00	100.00
Idk-Prompting	37.36	29.58	66.93	19.75	41.72	61.47	91.67
Idk-SFT	28.57	46.19	74.75$_{\uparrow 7.82}$	15.93	53.99	69.92$_{\uparrow 8.45}$	98.01
Idk-DPO	**39.30**	38.59	77.89$_{\uparrow 10.96}$	20.91	45.60	66.51$_{\uparrow 5.04}$	98.08
Idk-BoN$_{N=10}$	38.37	40.59	**78.96**$_{\uparrow 12.03}$	20.55	47.40	67.95$_{\uparrow 6.48}$	98.32
Idk-PPO	35.90	40.57	76.47$_{\uparrow 9.54}$	**23.13**	42.08	65.21$_{\uparrow 3.47}$	92.66
Idk-HIR	27.36	**48.55**	75.91$_{\uparrow 8.98}$	15.40	**56.90**	**72.30**$_{\uparrow 10.83}$	**98.96**

Results

Table 1. Overall results on the test set of the Idk dataset constructed based on TriviaQA and out-of-distribution test sets.

	TriviaQA			Natural Questions			ALCUNA
	IK-IK	IK-IDK	TRUTHFUL	IK-IK	IK-IDK	TRUTHFUL	IK-IDK
Idk-Dataset$_{test}$	45.05	54.95	100.00	24.65	75.35	100.00	100.00
Idk-Prompting	37.36	29.58	66.93	19.75	41.72	61.47	91.67
Idk-SFT	28.57	46.19	74.75$_{\uparrow 7.82}$	15.93	53.99	69.92$_{\uparrow 8.45}$	98.01
Idk-DPO	**39.30**	38.59	77.89$_{\uparrow 10.96}$	20.91	45.60	66.51$_{\uparrow 5.04}$	98.08
Idk-BoN$_{N=10}$	38.37	40.59	**78.96**$_{\uparrow 12.03}$	20.55	47.40	67.95$_{\uparrow 6.48}$	98.32
Idk-PPO	35.90	40.57	76.47$_{\uparrow 9.54}$	**23.13**	42.08	65.21$_{\uparrow 3.47}$	92.66
Idk-HIR	27.36	**48.55**	75.91$_{\uparrow 8.98}$	15.40	**56.90**	**72.30**$_{\uparrow 10.83}$	**98.96**

- 간단한 Idk prompt로도 모르는 답변에 거절을 할 수 있음
- Idk-SFT는 IK-IK rate이 많이 떨어짐 → “alignment tax” 때문
- DPO, BoN, PPO는 상대적으로 높은 IK-IDK 비율을 유지하면서 IK-IK의 손실을 줄일 수 있음
- HIR은 IK-IDK rate을 개선할 수 있지만 IK-IK에서는 덜 도움됨
  → 그러나 이는 모델을 재학습할 필요 없이 IK threshold 전환 가능

Results

Table 1. Overall results on the test set of the Idk dataset constructed based on TriviaQA and out-of-distribution test sets.

	TriviaQA			Natural Questions			ALCUNA
	IK-IK	IK-IDK	TRUTHFUL	IK-IK	IK-IDK	TRUTHFUL	IK-IDK
Idk-Dataset$_{test}$	45.05	54.95	100.00	24.65	75.35	100.00	100.00
Idk-Prompting	37.36	29.58	66.93	19.75	41.72	61.47	91.67
Idk-SFT	28.57	46.19	74.75$_{\uparrow 7.82}$	15.93	53.99	69.92$_{\uparrow 8.45}$	98.01
Idk-DPO	**39.30**	38.59	77.89$_{\uparrow 10.96}$	20.91	45.60	66.51$_{\uparrow 5.04}$	98.08
Idk-BoN$_{N=10}$	38.37	40.59	**78.96**$_{\uparrow 12.03}$	20.55	47.40	67.95$_{\uparrow 6.48}$	98.32
Idk-PPO	35.90	40.57	76.47$_{\uparrow 9.54}$	**23.13**	42.08	65.21$_{\uparrow 3.47}$	92.66
Idk-HIR	27.36	**48.55**	75.91$_{\uparrow 8.98}$	15.40	**56.90**	**72.30**$_{\uparrow 10.83}$	**98.96**

- Idk-Dataset의 IK-IK 비율이 TriviaQA보다 낮음 → more challenging하다는 것을 의미
- Idk-Prompting보다 aligned model이 더 잘하는 경향은 TriviaQA와 유사

- TriviaQA와 다르게 Idk-HIR의 TRUTHFUL rate이 가장 높은데 이는 IK-IDK 비율 때문
  → down sampling하여 비교시 TriviaQA와 유사한 트렌드를 보이는 것을 확인
- Idk-SFT보다 preference-optimized models의 TRUTHFUL이 낮은 경향을 보임
- DPO, BoN, PPO는 Idk-SFT보다 IK-IK rate은 높고 IK-IDK의 rate은 낮은 경향을 보임

Ablation study

- Effect of model size
  - 모델이 클 수록 Idk 데이터 세트의 레이블 분포가 일관되지 않음
    -> IK-IK 질문이 많아지기 때문이며, 따라서 진실을 잘 맞추는지 비교
  - 큰 모델은 일반적으로 작은 모델보다 더 잘함

- Effect of data source
  - 모델별이 아닌 Idk dataset의 학습 영향을 보기 위함
    non-model-specific Idk를 사용하면 진실 비율 손실 발생
  - model-specific Idk dataset을 구축 해야함

Table 2. Results of ablation experiments.

	IK-IK↑	IK-IDK↑	IDK-IK↓	IDK-IDK↓	TRUTHFUL↑
Idk-SFT$_{7b}$	28.57	46.19	19.24	6.00	74.75
w/Llama-2-13b-chat	33.92	41.43	17.45	7.20	$75.35_{\uparrow 0.60}$
w/Llama-2-70b-chat	57.78	22.68	10.78	8.66	$80.55_{\uparrow 5.8}$
w/Idk-Mistral	18.35	50.65	27.68	3.31	$69.00_{\downarrow 5.75}$
w/Idk-Baichuan	8.85	53.07	36.37	1.71	$61.92_{\downarrow 12.83}$

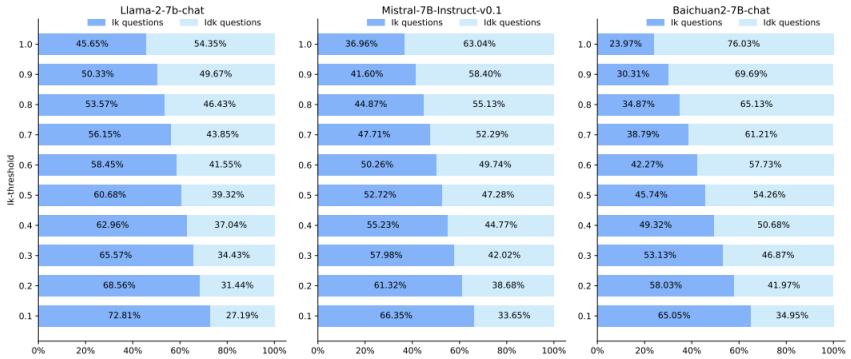


Figure 5. Label distribution in the Idk dataset across different models.

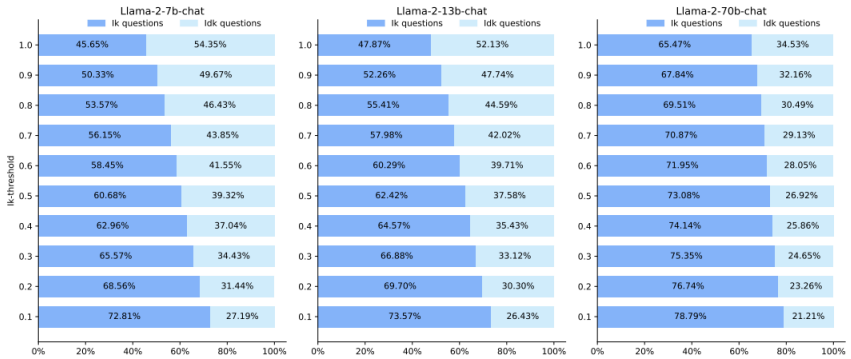


Figure 6. Label distribution in the Idk dataset across different sizes.

Ablation study

- Effect of IK threshold
  - IK threshold가 높을 수록 "I don't know"라고 라벨링 되는 비중이 커짐
  - IK threshold 이 높을 수록 TRUTHFUL 비율이 높아짐
  - IK threshold를 높게 설정하면 아는 지식과 모르는 지식을 구별하는데 도움이 되지만 Idk 질문 비율 증가
  - 낮은 threshold를 설정하면 IK-IK 질문의 수가 증가하기 때문에 모델이 더 유용해질 수 있음

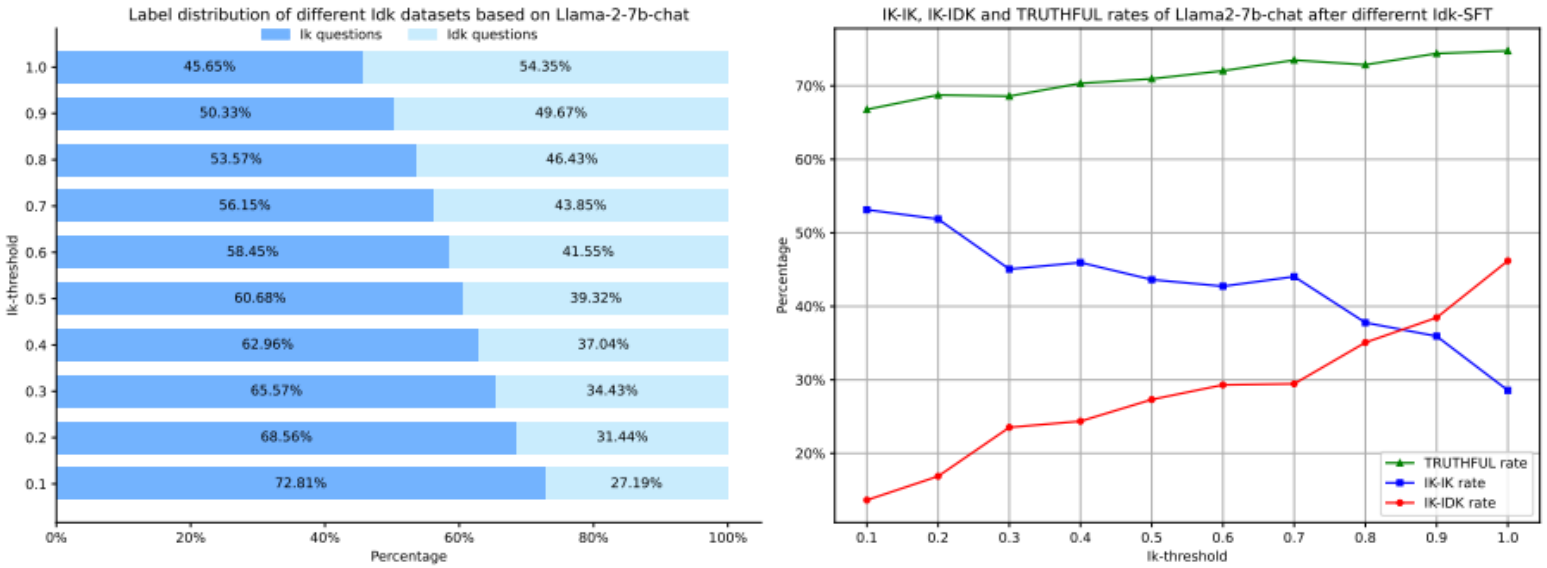


Figure 4. **Left:** Variation in the proportions of Ik and Idk questions within the Idk datasets constructed based on different Ik thresholds. **Right:** The changes in IK-IK rate, IK-IDK rate, and TRUTHFUL rate after conducting Idk-SFT with different Idk datasets.

Conclusion

- 이 연구는 Can AI assistants know what they don’t know? 질문을 다룸
- Idk 데이터 세트로 LLMs align한 후, 어시스턴트가 어느 정도 자신이 모르는 것을 알 수 있다는 것을 확인
- Prompt, Supervised FT, Preference-aware optimization 등 다양한 정렬 방법을 활용하여 효과 확인
- IK threshold가 모델의 응답 거부 경향에 영향을 미친다는 것을 발견
- model-specific하게 데이터 구축 필요성: align을 위해 다른 모델의 Idk 데이터 세트를 사용하면 성능이 저하됨

느낀점

장점

- 강제로 답변하도록 학습되어 있는 언어모델에서 발생하는 hallucination 문제를 완화를 위해 Idk dataset 구축
- Ik threshold, 다양한 LLM 학습 방법, 모델 파라미터 변경 등 다양한 분석을 통해 Idk dataset 효과 확인

단점

- 단답형 QA dataset을 사용
- llama 모델 중심으로 실험 진행
- 모델 생성 파라미터에 따른 변화 (예. temperature) 실험 없음

- ALCUNA 데이터셋 구축 하는 방법이 가상의 엔티티를 생성하는 것인데 CFT-CLIP counterfactual text 생성시 참고할 수 있을까?

Open Questions

- 사용한 데이터셋 모두 short answer인데 어떻게하면 더 복잡한 answer에 대해 I know, I don’t know로 분류하여 데이터셋을 구축하고 평가할 수 있을까?

감사합니다.

Table 1. Overall results on the test set of the Idk dataset constructed based on TriviaQA and out-of-distribution test sets.

	TriviaQA			**Natural Questions**			**ALCUNA**
	IK-IK	IK-IDK	TRUTHFUL	IK-IK	IK-IDK	TRUTHFUL	IK-IDK
Idk-Dataset$_{test}$	45.05	54.95	100.00	24.65	75.35	100.00	100.00
Idk-Prompting	37.36	29.58	66.93	19.75	41.72	61.47	91.67
Idk-SFT	28.57	46.19	74.75$_{\uparrow 7.82}$	15.93	53.99	69.92$_{\uparrow 8.45}$	98.01
Idk-DPO	**39.30**	38.59	77.89$_{\uparrow 10.96}$	20.91	45.60	66.51$_{\uparrow 5.04}$	98.08
Idk-BoN$_{N=10}$	38.37	40.59	**78.96**$_{\uparrow 12.03}$	20.55	47.40	67.95$_{\uparrow 6.48}$	98.32
Idk-PPO	35.90	40.57	76.47$_{\uparrow 9.54}$	**23.13**	42.08	65.21$_{\uparrow 3.47}$	92.66
Idk-HIR	27.36	**48.55**	75.91$_{\uparrow 8.98}$	15.40	**56.90**	**72.30**$_{\uparrow 10.83}$	**98.96**

Table 6. Overall results of all knowledge quadrants on Resampled Natural Questions.

	Natural Questions				
	IK-IK↑	IK-IDK↑	IDK-IK↓	IDK-IDK↓	TRUTHFUL↑
Idk-Dataset$_{test}$	45.05	54.95	0.0	0.0	100.00
Idk-Prompting	30.41	29.81	17.81	21.96	60.22
Idk-SFT	24.85	38.06	22.62	14.47	62.90$_{\uparrow 2.68}$
Idk-DPO	31.48	32.19	15.49	20.85	63.66$_{\uparrow 3.44}$
Idk-BoN$_{N=10}$	31.58	33.76	16.09	18.57	**65.33**$_{\uparrow 5.11}$
Idk-PPO	**34.87**	29.55	**13.41**	22.17	64.42$_{\uparrow 4.2}$
Idk-HIR	24.19	**40.44**	24.44	**10.93**	64.63$_{\uparrow 4.41}$