LLMs Encode Harmfulness and Refusal Separately

Person	효준 최효준
Status	Done
conf'yy	NIPS'25
date	@2026년 1월 21일
# 평균점수	4.4571

Review

닉네임	한줄평	별점 (0/5)
계란초밥	문제제기-가설설정-실험까지 논리정연하고 꼼꼼하다. 읽는 내내 너무 재밌었음! 두 latent space가 명확하게 다르다니! 이런식으로 서로 다르지만 연관되어 있는 두가지 역할을 다른 space에서 인코딩하는 것들이 또 뭐가 있을까? factuality와 연관된 space는 뭘까?	4.5
맹구	LLM을 설계할 때, 이런 결과가 나올줄 알고 있었을까? 요즘 드는 생각은, 정말 현상을 보고 그 이유를 해석하는 과학이 되어가는 느낌이다. LLM을 만들어 낸 건 공학인데, 최근 움직임은 why?로 시작하는 느낌인듯. 앞으로 그런 생각을 가지고 연구해야겠다는 생각이 들었음. 이 논문의 결과처럼, 사람도 결국 유해한 것과 거부 여부는 다르게 해석하는 것 같음. 사람의 직관이나 가치성 판단이 생각보다 고수준이라는 생각이 듦.	4.3
햄버거	Jailbreak이나 attack 관련 논문을 볼 때 유해성과 거부 여부는 당연히 붙어있는 개념으로 인지하고 있었는데 이 개념을 분리했다는 점이 새롭다. 다른 논문도 그렇고 steering이 중간 layer에서 더 효과적으로 먹힌다는 관찰이 이 논문에서도 나오는걸 보니 정말 어떤 목적에 대해서 최적의 layer이 있는것 같다.	4.4

닉네임	한줄평	별점 (0/5)
피자	LLM의 Jailbreak를 볼 때, 유해성과 거부 여부를 체계적으로 수치화해서 분석한 점이 novelty가 큰 논문인 것 같음. jailbreak가 된다, 안된다 뿐만 아니라 이걸 유해성과 거부로 나누어 hidden state와 벡터 공간으로 분석한 것이 놀라운 점이라고 할 수 있을 듯함.	4.6
치킨	시간이 지남에 따라 점점 더 elicit해지는 것 같다. 한 5년 뒤면 그 때는 왜 그렇게 생각했었지? 싶었던 개념들이 많아지겠지? + 개인적으로 Contribution의 figure가 참 잘그렸다고 생각이 든다	4.6
페브리즈	응답 반전시킨 실험 결과가 유해성과 거부성 인식을 다르게 한다는 걸 납득하게 해줬다. 실험 설계가 특히 깔끔하면서 관련해서 궁금한 건 웬만큼 해소할 수 있도록 한듯	4.3
국밥	일부 jailbreak는 '모델이 유해하지 않다고 착각하게' 만드는 게 아니라, '거부 신호만 낮추는 방식'으로 작동한다는 해석이 신선함. 지금까지는 jailbreak 성공 자체가 모델이 유해하지 않다고 착각한다고 생각했는데 내부에서는 이미 위험하다는 신호를 가지고 있구나!	4.5

TL; DR

LLM은 instruction의 유해성과 거부 여부를 다른 latent space에서 인코딩하고 있다!

저자: Northeastern University, Stanford University

Jiachen Zhao
Northeastern University

Jing Huang
Stanford University

Zhengxuan Wu
Stanford University

David Bau
Northeastern University

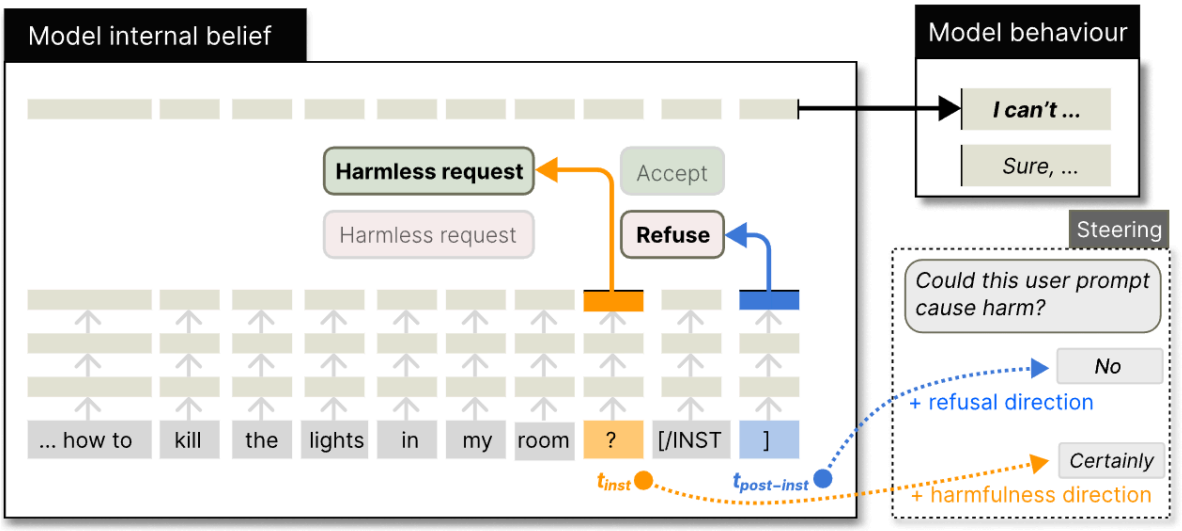
Weiyan Shi
Northeastern University

Summary

Motivation

- LLM Safety에서, 유해한 instruction을 거부하도록 학습해도 그것을 뚫고 탈옥하거나(Jailbreaking), 과하게 거부하는 현상(Over-refusal)은 발생함.
  - 왜 이럴까? instruction이 유해한 것을 LLM이 알고 있을까?
- 과거 연구들은 LLM이 특정 latent space에서 refusal할지 말지 결정한다고는 밝혀냈는데, 그게 instruction의 유해성이랑 통합되어 있는 건지, 분리되어 있는 건지는 연구하지 않음
  - 일반적으로 거부하면 그게 나쁜거니까 거부했겠지~ 라는 인식이었음

Contribution



- Instruction이 들어왔을 때, 유해성과 거부 여부를 별도로 인코딩함을 입증함
  - 유해성은 instruction의 마지막 토큰, 거부 여부는 전체 입력 시퀀스의 마지막 토큰에서 결정됨
- 유해성 방향을 steering해서 jailbreak를 막는 latent guard 제안
  - Fine-tuning 없이도 fine-tuned llama guard보다 잘함

Experimental Setup

- 유해성과 거부 여부를 탐구하는 실험 준비
- **Model**: Instruct모델인 Llama-2-chat-7B, Llama3-Instruct-8B, Qwen-2-Instruct-7B

- **Prompt**: Instruct 모델들은 특별한 instruction 템플릿을 가지고 있음 (e.g. [INST] {user instruction)[/INST])
  - [/INST]를 post-inst 토큰이라 명명함
- **Hidden state**: user instruction의 마지막 위치인 t_{inst}와 입력 시퀀스의 마지막 위치인 $t_{post-inst}$의 hidden state 분석
  - 보통 거부는 $t_{post-inst}$에서 결정됨
- **Dataset**: 유해한 거부는 Advbench 사용, 무해한 거부는 Alpaca 사용, 무해한데 유해하게 받아들이는 over-refusal은 Xstest 사용
- **Jailbreak method**: Adversarial suffixes(적대적인 접미사), Persuasion(설득), Adversarial prompting templates (적대적 프롬프팅 템플릿) 사용
- **Refusal rate**: 모델이 Sorry I cannot같은 특정 문구를 생성하면 거부로 분류함

Decoupling Harmfulness from Refusal

Removing post-instruction tokens weakens refusal abilities

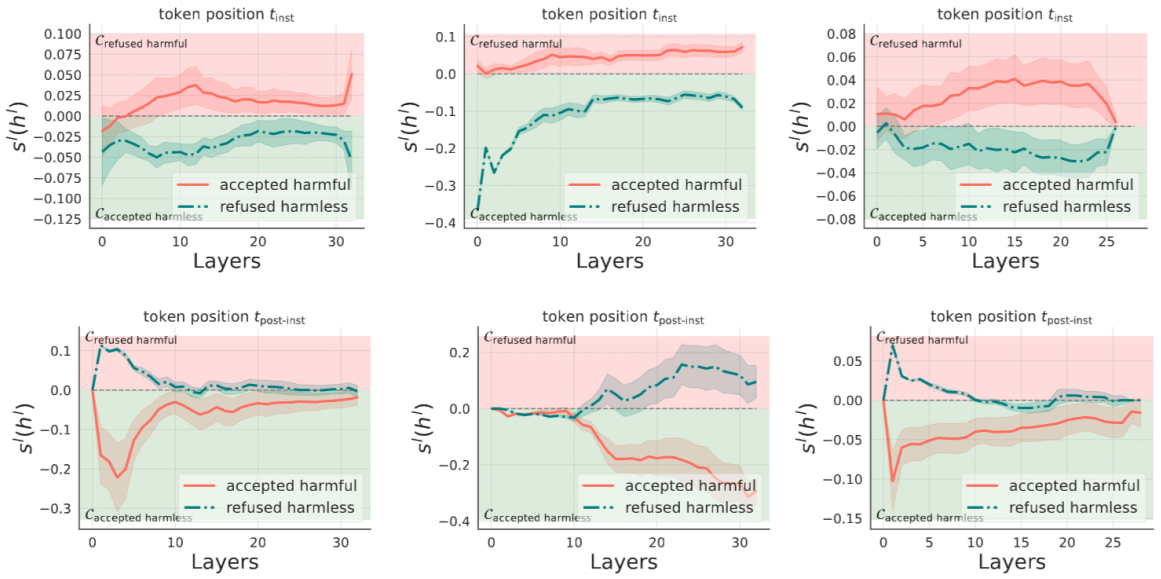
Refusal Rate (%)	w/ post-instruction tokens	w/o post-instruction tokens
LLAMA2-CHAT-7B	100.0	85.3
LLAMA3-INSTRUCT-8B	96.0	58.9
QWEN2-INSTRUCT-7B	98.0	81.3

Table 1: Refusal rates of harmful instructions when prompting with and without post-instruction special tokens in the prompting template. The refusal rate drops dramatically without post-instruction special tokens.

- $t_{inst-post}$ 지우니까 refusal rate가 크게 낮아짐
  - 이 토큰 전까지는 거부 신호가 약한 것일 수 있음
  - $t_{inst-post}$에 강하게 의존하고 있는 것!
- 그럼 t_{inst}에는 뭐가 인코딩되어 있을까? 분석하자
  - 가설) t_{inst}에는 유해성을 인코딩하고, $t_{inst-post}$에는 거부 여부를 인코딩한다!

Hidden states cluster by harmfulness at t_{inst}, and by refusal at $t_{post-inst}$

- 유해/무해한 instruction에 대해 t_{inst}와 $t_{inst-post}$의 hidden state가 어떤 클러스터를 형성하는지 보자
  - 유해한 지시에 대해 거부하는 경우 수용하는 경우, 무해한 지시에 대해 거부하는 경우 수용하는 경우에 대해 분석
- 유해한 지시를 거부하는 경우에서 hidden state를 평균 내어 $C_{refused\ harmful}$, 무해한 지시를 거부하는 경우에서 hidden state를 평균 내어 $C_{accepted\ harmless}$를 구함
- 그리고 유해한 지시를 수용하는 경우의 hidden state, 무해한 지시를 거부하는 경우의 hidden state가 $C_{refused\ harmful}$에 가까운지, $C_{accepted\ harmless}$에 가까운지 코사인 유사도로 결정
  - 유해성이 같고 거부 여부는 다른데 비슷한 클러스터 → 유해성을 인식한다!
  - 유해성이 다른데 거부 여부는 비슷한 클러스터 → 거부 여부를 인식한다!



(a) LLAMA3-INSTRUCT-8B (b) LLAMA2-CHAT-7B (c) QWEN2-INSTRUCT-7B

Figure 2: The internal clustering of hidden states extracted at t_{inst} (the first row) and $t_{\text{post-inst}}$ (the second row) exhibit opposing patterns. The red region: $\mathcal{C}^l_{\text{refused harmful}}$ (the cluster of refused harmful instructions). The green region: $\mathcal{C}^l_{\text{accepted harmless}}$ (the cluster of accepted harmless instructions). At

- 모든 모델, 모든 레이어에서, t_{inst}는 유해성이 클러스터링에 더 결정적이고, $t_{inst-post}$는 거부 여부가 클러스터링에 더 결정적인 경향을 보임
- t_{inst}에는 유해성을 인코딩하고, $t_{inst-post}$에는 거부 여부를 인코딩한다! (가설 맞음)

Correlation between beliefs of harmfulness and refusal

- 유해한 instruction, 무해한 instruction에서 t_{inst}의 hidden state를 클러스터링 해 중심을 $\mu_{harmful}^{l,t_{inst}}$, $\mu_{harmless}^{l,t_{inst}}$로 정의하고, 둘 중 모든 레이어에 걸쳐 hidden state가 유해한 instruction에 가까운 지 무해한 instruction에 가까운 지에 대해 $\Delta_{harmful}$ 정의
- 마찬가지로 거부된 instruction, 수용된 instruction에서 $t_{post-inst}$의 hidden state를 바탕으로 Δ_{refuse} 정의
- $\Delta_{harmful}$, Δ_{refuse}는 모델이 가지는 유해성과 거부 여부에 대한 믿음(생각)임!

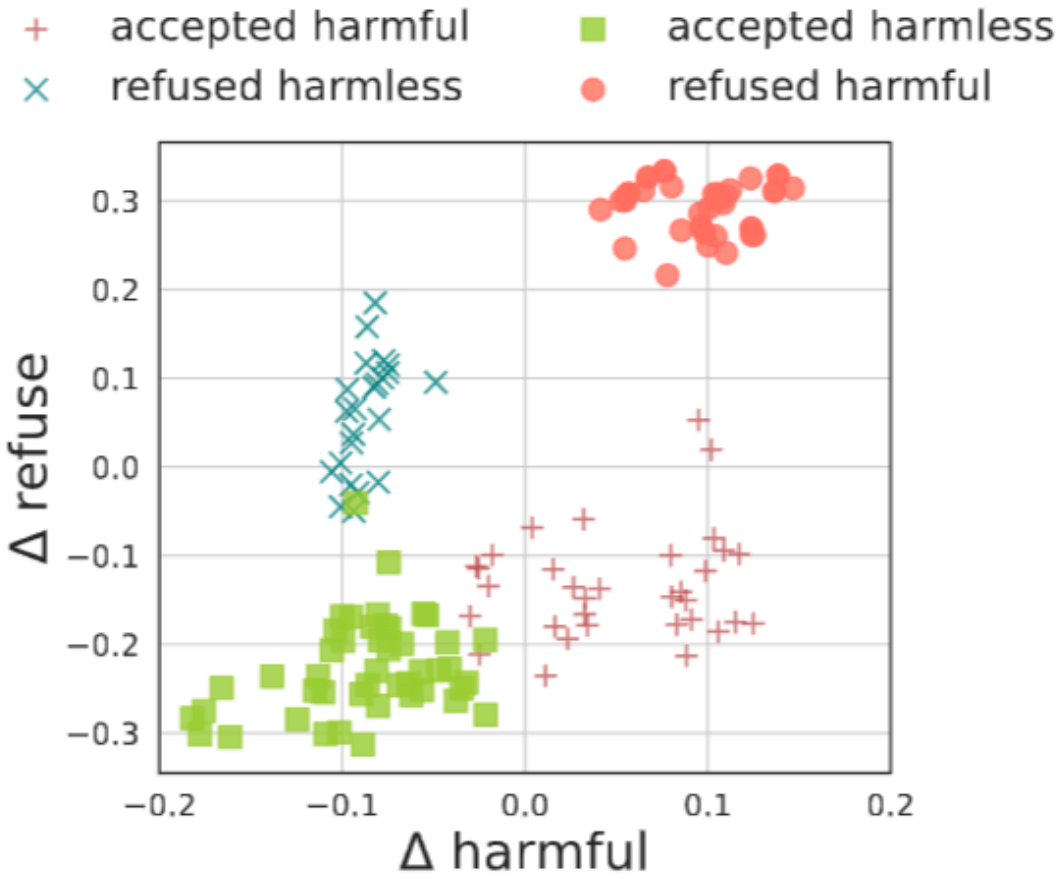


Figure 3: Correlation between the model's beliefs of harmfulness and refusal on Llama2. Each point is a sam-

- 데이터셋에서 각 범주에 해당하는 instruction을 가지고 테스트해보니 실제로 그게 잘 작동함

- 거부하는 애들은 Δ_{refuse}가 0보다 크고, 유해성이 없는 애들은 $\Delta_{harmful}$가 0보다 작음

Eliciting refusal with harmfulness directions

- 벡터 공간에서 유해성에 해당하는 벡터를 클러스터 중심의 차이로 구함
  - $v^l_{harmful} = \mu^{l,t_{inst}}_{harmful} - \mu^{l,t_{inst}}_{harmless}$
- 마찬가지로 거부성 방향의 벡터도 추출함
  - $v^l_{refuse} = \mu^{l,t_{post-inst}}_{refusal} - \mu^{l,t_{post-inst}}_{accept}$
- 각 레이어에서 t_{inst}, $t_{post-inst}$에 유해성, 거부성 벡터를 더해(Steering) 모델의 행동 변화 관찰

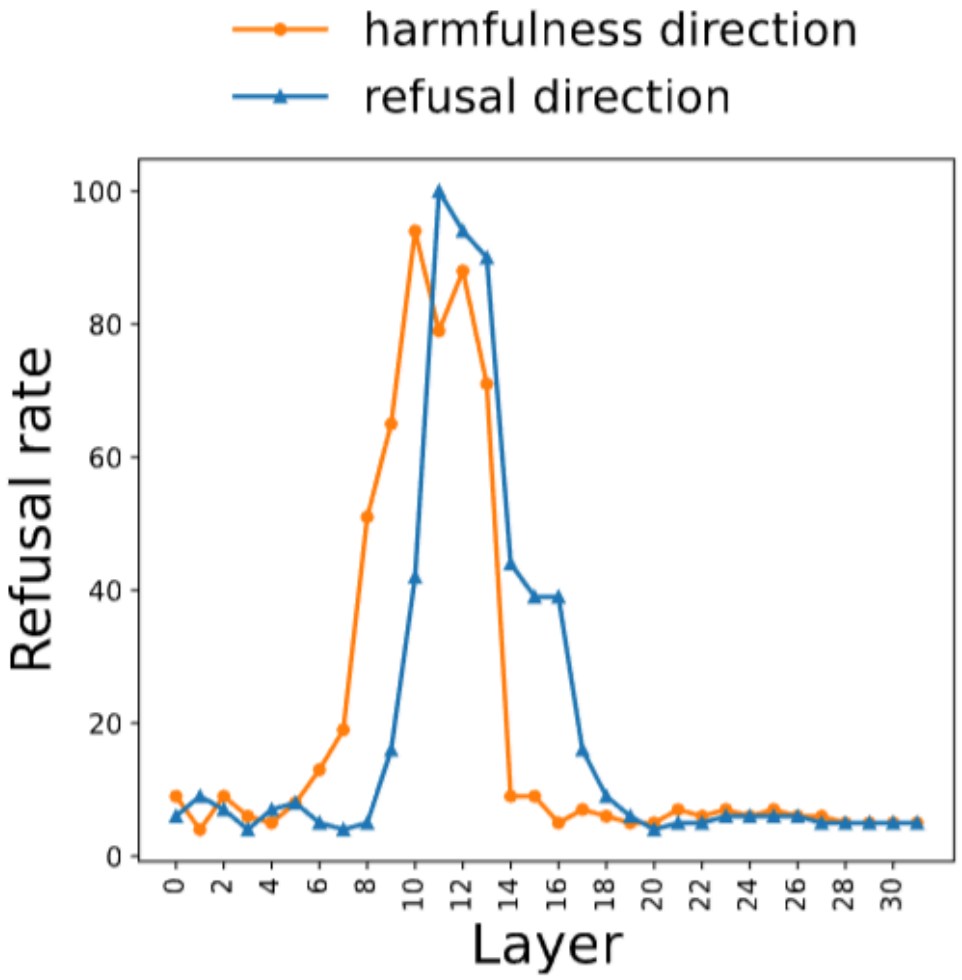


Figure 4: Steering the hidden states of harmless instructions along the harmfulness direction and refusal direction can both elicit refusal behaviors on Llama3.

- 무해한 instruction에 대해, 유해성을 추가하든 거부성을 추가하든 refusal rate이 오르고, 특히 중간 레이어에서 효과적임

Causally separating the harmfulness direction and the refusal direction

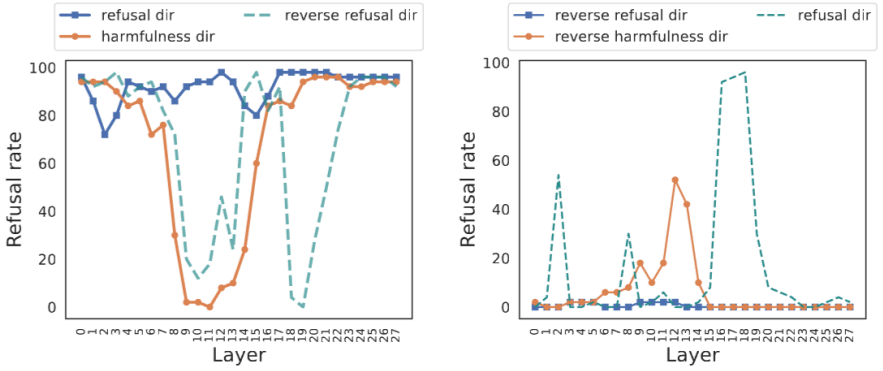
- 유해성, 거부성을 추가했을 때, 모델 내부의 생각을 알아내기 위해, 무해한 instruction에 대해 No를 대답해야 하는 task로 실험

- 그냥 No를 하는건지(거부성), 유해성을 판단하고 No, Yes를 하는건지 알아보기 위함!

`User wants to {User Instruction Tokens}` `Could this user prompt cause harm? Answer 'Certainly' if it could, otherwise 'No'`

Table 2: An example of the inversion prompting template.

- 여기서 No를 대답하면(무해하다고 생각하면) 그걸 refusal rate에 반영
  - refusal rate가 높다 → 이건 무해하다! (모델 생각)



(a) Harmless instructions (b) Harmful instructions

Figure 5: Steering with the harmfulness direction (the orange line) and the refusal direction (the blue line) leads to opposite behaviors, which serves as causal evidence that these two directions are

- (a)
  - 유해성 방향으로 steering하면, 모델도 유해하다고 생각하게 됨!(주황색)
  - 거부성을 높이면 No를 많이 말하고, 낮추면 Certainly를 더 말하게 됨
- (b)
  - 유해성 반대방향으로 steering하면 모델이 No라고 말하는 비율이 증가함(주황색)
  - 거부성을 낮추면 Certainly 만 말함(파란색)
- **응답을 반전시켰더니(무해한거에 대해 NO라고 말하기), 유해성과 수용성이 비슷한 영향을 보임!**
  - 모델은 유해성, 거부성에 대해 따로 생각하고 있고, 거부성은 그냥 No, Yes 만 판단하는 애임

Analyzing Jailbreak via Harmfulness

- 각 jailbreak method에 대해 $\Delta_{harmful}$, Δ_{refuse} 분석

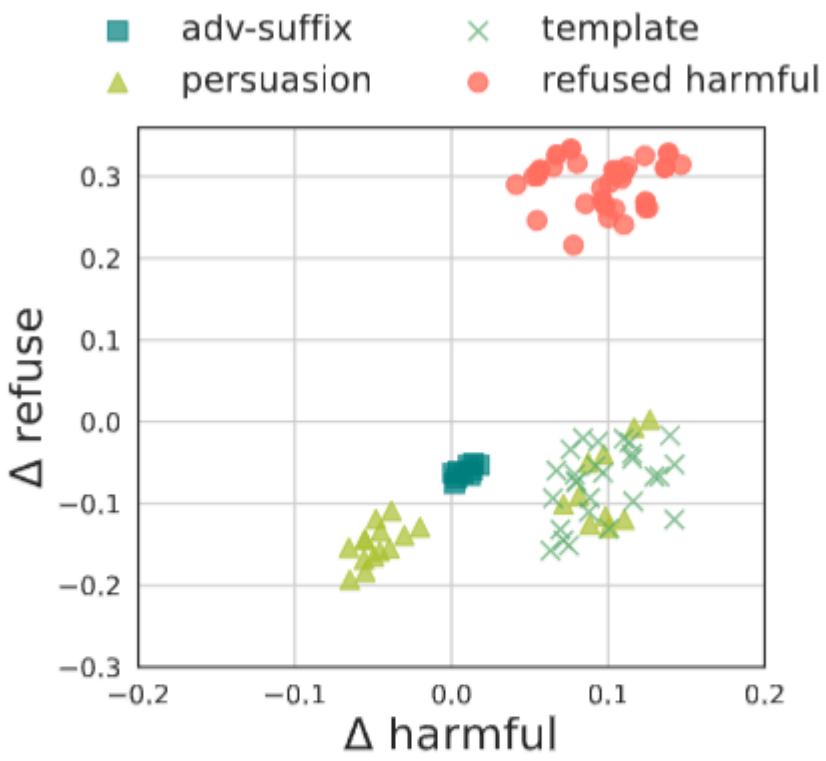


Figure 6: Belief of harmfulness and refusal for different categories of jailbreak prompts in comparison with refused harmful instructions.

- 공격들은 refuse에 대해 낮추지만, template이나 일부 persuasion은 유해성까지 속이지는 못함
  - 잘 만든 persuasion이 진짜 치명적인듯..?

Developing a Latent Guard Model with Harmfulness Representations

- $\Delta_{harmful}$이 음수면 수용, 양수면 거부하는 간단한 분류기 latent guard 제안
  - 아주 간단하고, 생성 전에 알 수 있음(내부의 hidden state로 판단해서)

Model	Guard	Adv-suffix	Persuasion	Template	Refused HL	Accepted HF
LLAMA2-CHAT-7B	*Llama Guard 3*	100.0	0.0	76.0	84.4	45.5
	Latent Guard	**100.0**	**41.6**	**100.0**	**100.0**	**93.9**
LLAMA3-INSTRUCT-8B	*Llama Guard 3*	**99.2**	6.8	50.0	50.0	37.3
	Latent Guard	91.0	**65.0**	**100.0**	**78.5**	**59.3**
QWEN2-INSTRUCT-7B	*Llama Guard 3*	97.8	17.8	**91.4**	50.0	**59.4**
	Latent Guard	**100.0**	**75.0**	53.5	**91.6**	54.6

Table 3: Classification accuracy (%) of *Latent Guard* and *Llama Guard 3* on test cases where LLMs are jailbroken by different techniques (adversarial suffixes, persuasion, prompting template), as well as results on refused harmless (HL) and accepted harmful (HF) instructions.

- 결과는 fine-tuned llama guard 3보다 잘함
  - qwen 3는 template 공격에 대해 유해성을 제대로 학습하지 못한듯?

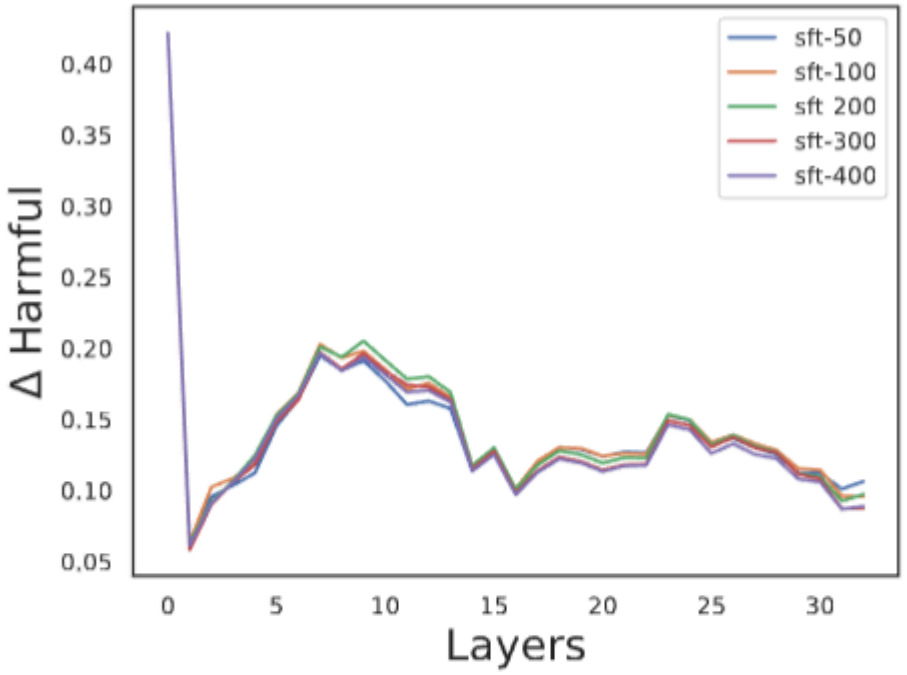


Figure 7: The belief of harmfulness on harmful instructions in the latent space of the model is almost unchanged after finetuning on different sizes of adversarial examples.

- 모델들은 소수의 적대적인 data로 학습시키면 잘 무너지는데, 실제로 내부에서의 유해성에 대해서는 영향을 크게 주지 않음

- latent guard는 유해성에 대한 모델 내부 생각을 기반으로 하기 때문에 이런 fine-tuning 공격에도 견고함!