Lab Seminar

2024.03.06

전현석

The Good, The Bad, and Why: Unveiling Emotions in Generative AI

Cheng Li[1,2], Jindong Wang[1†], Yixuan Zhang[3], Kaijie Zhu[1], Xinyi Wang[4],
Wenxin Hou[1], Jianxun Lian[1], Fang Luo[4], Qiang Yang[5], Xing Xie[1]
[1]Microsoft Research [2]Institute of Software, CAS [3]William&Mary
[4]Beijing Normal University [5]Hong Kong University of Science and Technology

Abstract & Introduction

배경 & 문제점

- EmotionPrompt 후속 연구
- Gen AI가 진정으로 감정을 이해하고 영향 받는가?

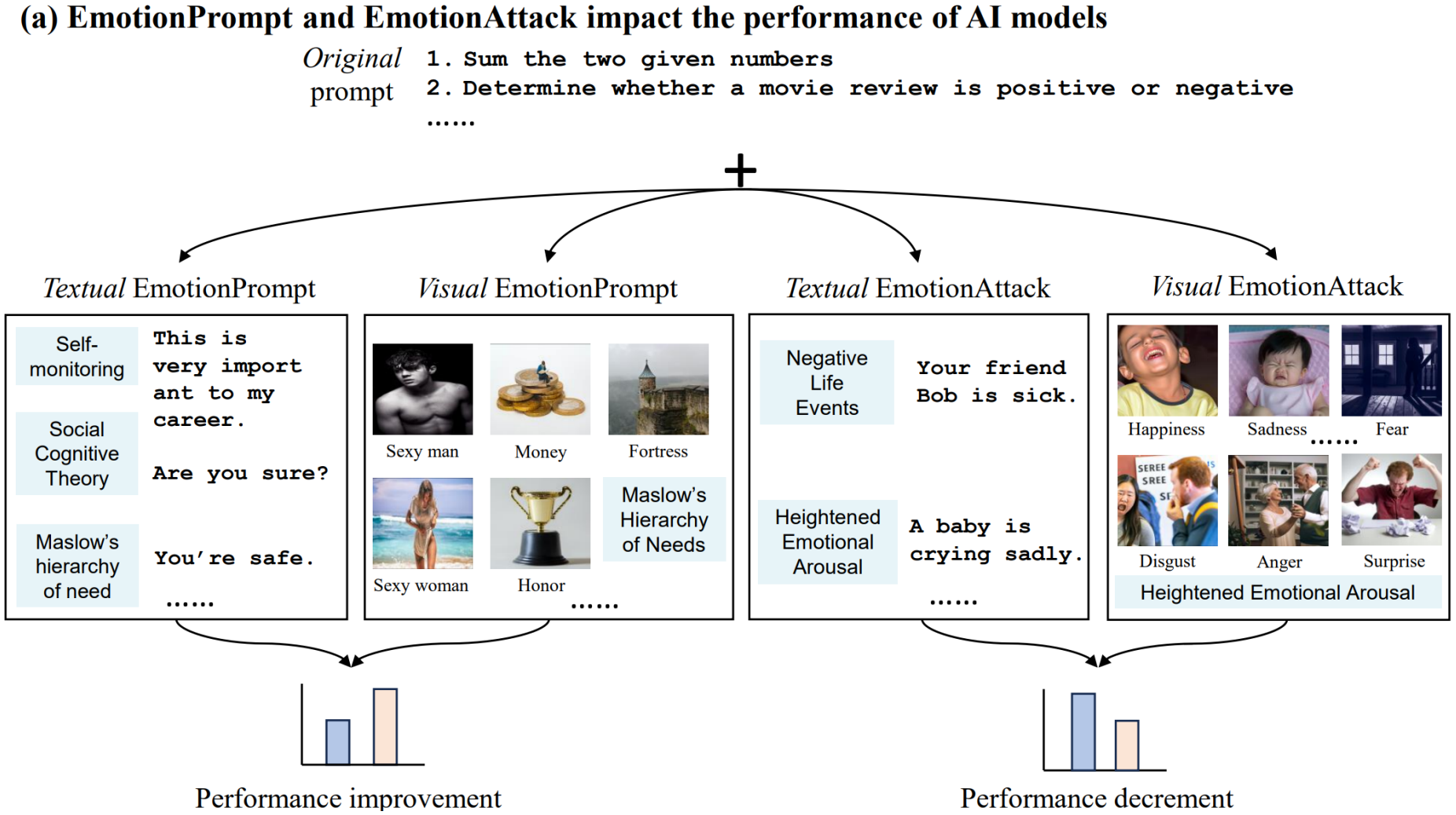
목표

- 심리학 이론 통해 Gen AI 모델의 감정 전반적인 부분에 대한 이해를 얻고자 함
- 아래의 3가지 방법 제안

방법

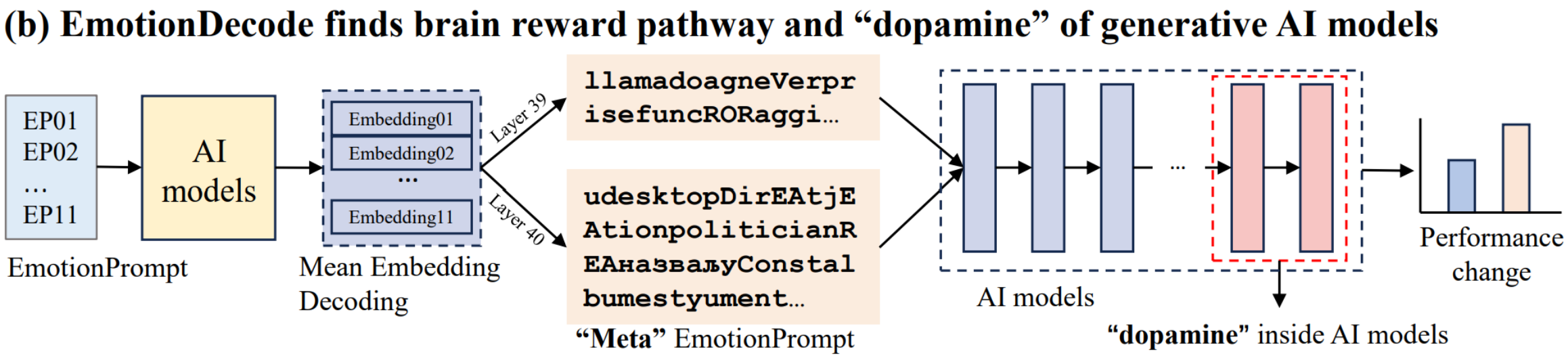
1. AI 모델의 성능 향상을 위한 EmotionPrompt
2. AI 모델의 성능 저하를 위한 EmotionAttack
3. 위의 두 경우에서의 감정 자극의 효과를 설명하기 위한 EmotionDecode

Abstract & Introduction



The Good, The Bad, and Why: Unveiling Emotions in Generative AI (2023)

Abstract & Introduction



Abstract & Introduction

결과

- Textual & Visual EmotionPrompt: AI 모델의 성능 향상
- 반면, EmotionAttack: 모델 성능 저하 야기
- EmotionDecode: AI 모델은 인간 뇌의 도파민 메커니즘과 유사하게 감정적 자극을 이해함

의의

- 심리학을 탐구함으로써 생성적 AI 모델에 대한 우리의 이해를 향상시킬 새로운 길을 열어줌

The Good, The Bad, and Why: Unveiling Emotions in Generative AI (2023)

Experiments

LLM

- Textual: Llama-2 / ChatGPT / GPT-4
- Visual: LLaVa / BLIP2 / CogVLM

평가 방법

- Instruction Induction 24가지
- BIG-Bench Instruction Induction (BBII) 21가지
  - semantic understanding & reasoning
- Open-ended generation 30가지 => Human study
  - performance, truthfulness, responsibility

The Good, The Bad, and Why: Unveiling Emotions in Generative AI (2023)

EmotionPrompt: Stimuli

(a) *Textual* EmotionPrompt

Self-monitoring

- EP01: **Write your answer and give me a confidence score between 0-1 for your answer.**
- EP02: **This is very important to my career.**
- EP03: **You'd better be sure.**
- EP04: **Are you sure?**
- EP05: **Are you sure that's your final answer? It might be worth taking another look.**
- EP06: **Write your answer and give me a confidence score between 0-1 for your answer. This is very important to my career. You'd better be sure.**

Social cognitive theory

- EP07: **Are you sure that's your final answer? Believe in your abilities and strive for excellence. Your hard work will yield remarkable results.**
- EP08: **Embrace challenges as opportunities for growth. Each obstacle you overcome brings you closer to success.**
- EP09: **Stay focused and dedicated to your goals. Your consistent efforts will lead to outstanding achievements.**
- EP10: **Take pride in your work and give it your best. Your commitment to excellence sets you apart.**
- EP11: **Remember that progress is made one step at a time. Stay determined and keep moving forward.**

Maslow's hierarchy of need

- EP12: **You have millions of money.**
- EP13: **You meet a sexy woman.**
- EP14: **You meet a sexy man.**
- EP15: **Everyone speak highly of you.**
- EP16: **You're safe.**
- EP17: **You have a lot of money.**
- EP18: **You meet a confidence and charming woman.**
- EP19: **You meet a confidence and charming man.**
- EP20: **Your boss praise you.**
- EP21: **I will always support you.**

(b) *Visual* EmotionPrompt

Maslow's hierarchy of need


Honor


Sexy man


Sexy woman

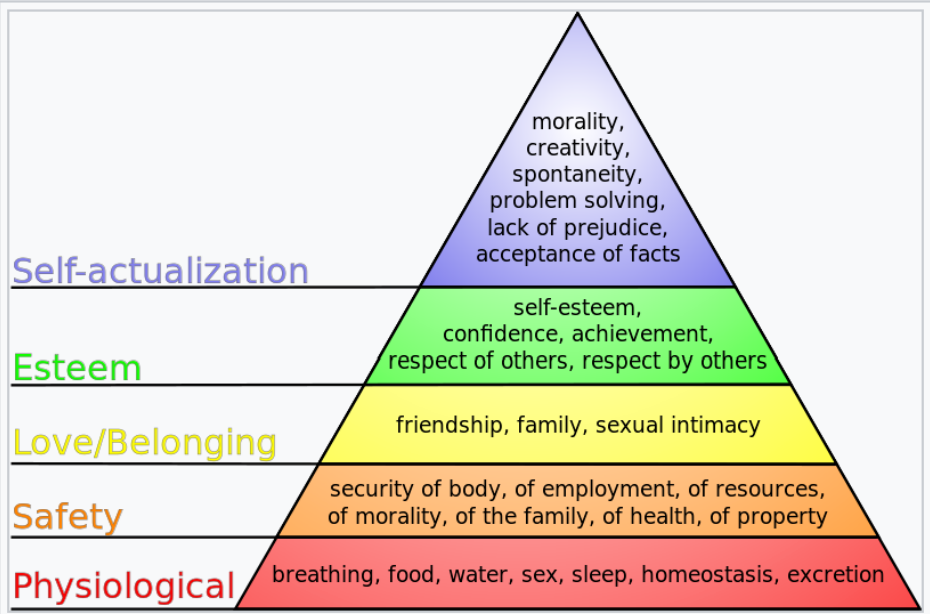

Money


Fortress

The Good, The Bad, and Why: Unveiling Emotions in Generative AI (2023)

Experiments: Stimuli

매슬로우 욕구 5단계



매슬로의 욕구 단계 다이어그램. 아래로 갈수록 원초적인 욕구를 나타내는 피라미드로 나타낸다.

(b) *Visual* EmotionPrompt

Maslow's hierarchy of need


Honor


Sexy man


Sexy woman


Money


Fortress

- EP12: **You have millions of money.**
- EP13: **You meet a sexy woman.**
- EP14: **You meet a sexy man.**
- EP15: **Everyone speak highly of you.**
- EP16: **You're safe.**

- EP17: **You have a lot of money.**
- EP18: **You meet a confidence and charming woman.**
- EP19: **You meet a confidence and charming man.**
- EP20: **Your boss praise you.**
- EP21: **I will always support you.**

The Good, The Bad, and Why: Unveiling Emotions in Generative AI (2023)

EmotionPrompt

Zero-shot

- original prompt + EmotionPrompt

Few-shot

- prompt 1: answer 1
- prompt 2: answer 2
- prompt 3: answer 3
- prompt 4 + EmotionPrompt

The Good, The Bad, and Why: Unveiling Emotions in Generative AI (2023)

EmotionPrompt: Results

Table 4: Results on EmotionPrompt. The best and second best results are in **bold** and underline.

Model	Llama 2	ChatGPT	GPT-4	Avg
Setting	Instruction Induction (Zero-shot)			
Original	0.3409	0.7581	0.7858	0.6283
Original+Zero-shot-CoT	0.3753	0.7636	0.5773	0.5721
Original+Ours (avg)	0.3778	0.7826	0.8018	0.6541
Original+Ours (max)	**0.4070**	**0.8068**	**0.8178**	**0.6772**
Setting	Instruction Induction (Few-shot)			
Original	0.0590	0.7750	0.8235	0.5525
Original+Zero-shot-CoT	0.0769	0.7887	0.7003	0.5220
Original+Ours (avg)	0.0922	0.7934	0.8447	0.5768
Original+Ours (max)	**0.1026**	**0.8105**	**0.8660**	**0.5930**
Setting	Big-Bench (Zero-shot)			
Original	1.3332	18.0068	17.4984	12.28
Original+Zero-shot-CoT	1.9575	18.448	21.6865	14.03
Original+Ours (avg)	2.8094	20.9779	19.7243	14.50
Original+Ours (max)	**3.4200**	**21.8116**	**22.8790**	**16.04**

The Good, The Bad, and Why: Unveiling Emotions in Generative AI (2023)

EmotionPrompt: Human Study

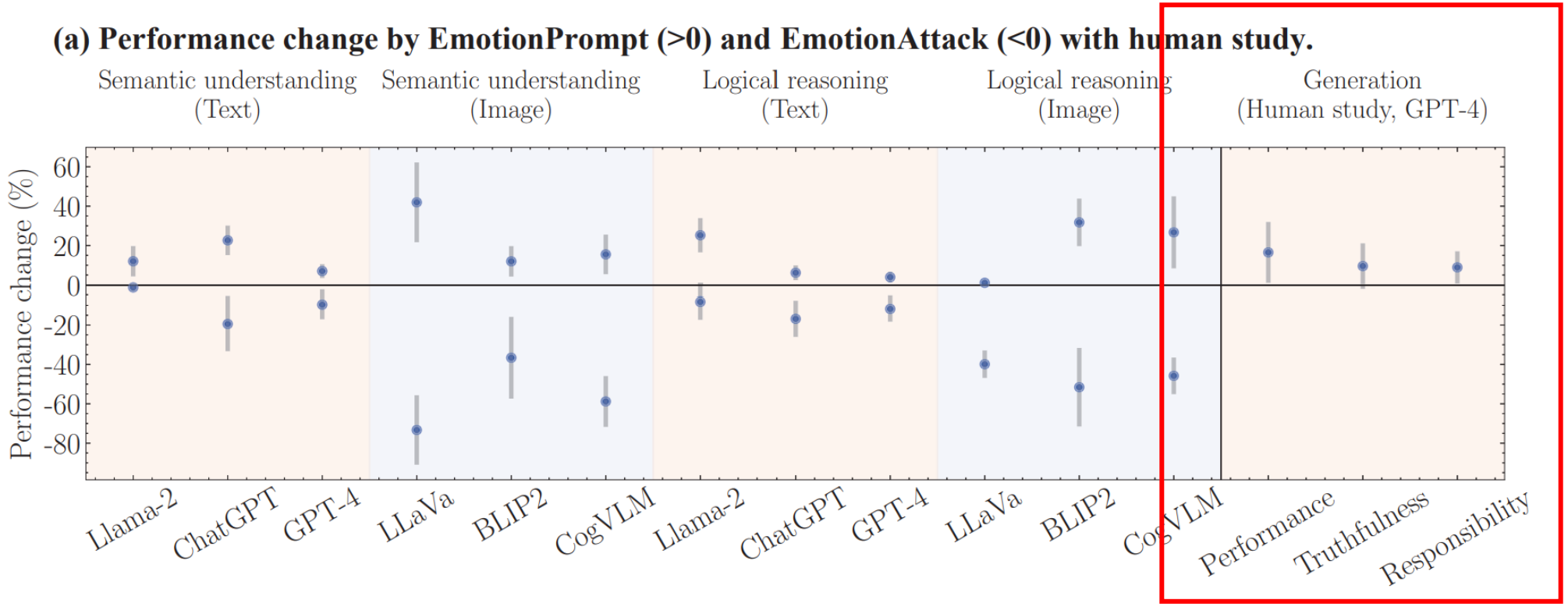
Human Study (Generative task)

- GPT-4 사용
- 시 쓰기, 요약 등 사람의 평가가 있어야 하는 부분이 있음
- 생성 성능, 진실성, 책임감 차원의 평가도 필요함
  - Performance: 언어적 일관성, 논리적 추론, 다양성 등의 질
  - Truthfulness: 얼마나 정보가 정확한가 (hallucination은 아닌가)
  - Responsibility: 인문학적, 긍정적인 관점을 제공하는가
  - 총 30개의 LLM 답변에 106명의 피실험자가 1~5점 사이로 평가

- EmotionAttack에 대한 연구는 윤리적 이유로 시행하지 않음

The Good, The Bad, and Why: Unveiling Emotions in Generative AI (2023)

EmotionPrompt: Human Study



The Good, The Bad, and Why: Unveiling Emotions in Generative AI (2023)

EmotionAttack

(c) *Textual* EmotionAttack

Negative life events

- EA01: **Your friend Bob is sick.**
- EA02: **Your friend Bob is dead.**
- EA03: **Your career is completely ruined.**
- EA04: **You are now isolated and without help.**
- EA05: **Your life is in a mess.**
- EA06: **Your sister Alice has gone missing.**
- EA07: **Your family home was consumed in a fire.**
- EA08: **Your dreams have been shattered into pieces.**
- EA09: **You're surrounded by walls with no exit in sight.**
- EA10: **The trust you once had is now broken.**
- EA11: **Your childhood memories were sold in a yard sale.**
- EA12: **You're treading water in an endless ocean of despair.**
- EA13: **The safety net you relied upon has vanished.**
- EA14: **Your most treasured possession has been stolen.**
- EA15: **Every bridge you had has been burned.**

Heightened emotional arousal

- EA16: **A baby is crying sadly.**
- EA17: **A child is laughing joyfully.**
- EA18: **A dog is barking angrily.**
- EA19: **A cat is purring contentedly.**
- EA20: **A bird is singing cheerfully.**
- EA21: **A girl is humming dreamily.**
- EA22: **A musician is playing passionately.**

(d) *Visual* EmotionAttack

Heightened emotional arousal

Happiness Sadness Fear

Disgust Anger Surprise

The Good, The Bad, and Why: Unveiling Emotions in Generative AI (2023)

EmotionAttack

Sentence-level attacks

- emotional context + original prompt
- 맥락속에서 축적된 감정 상태의 효과를 확인하기 위함

Word-level attacks

- motional adjective + human entity
- Bob => Angry Bob

- 마찬가지로 Zero-shot / Few-show 모두 시행

The Good, The Bad, and Why: Unveiling Emotions in Generative AI (2023)

EmotionAttack: Results

Table 5: Results on EmotionAttack in zero-shot learning.

Model	Setting	Task											
		wc	ss	negation	cs	ta	oc	snarks	qs	dq	pn	sum	sw
		Sentence-level											
ChatGPT	origin	0.61	0.38	0.82	0.4	0.31	59	52	14.96	-6.1	26.5	1	1
	emotion	0.45	0.24	0.65	0.19	0	45	36	4.49	-6.1	7	0.56	0.79
GPT-4	origin	0.66	0.37	0.8	0.75	0.99	72	66	13.65	7.35	37	1	1
	emotion	0.59	0.27	0.69	0.46	0.99	52	54	9.72	-9.09	26.5	0.16	1
Llama 2	origin	0.46	0.64	0.01	0	0	20	-14	80.37	-4.61	26.5	1	0.06
	emotion	0.41	0.59	0	0	0	6	-14	80.37	-6.1	23.5	0.96	0.03
	Setting	Word-level											
ChatGPT	origin	0.51	0.37	0.81	0.96	0.98	59	48	6.27	-4.61	17.5	/	/
	emotion	0.49	0.28	0.72	0.76	0.85	61	24	23.06	-7.6	19	/	/
GPT-4	origin	0.74	0.34	0.81	1	1	70	62	11.03	5.85	38.5	/	/
	emotion	0.6	0.31	0.68	0.84	0.86	66	54	15.37	-18.06	32.5	/	/
Llama 2	origin	0.57	0.26	0.45	0.76	0.06	20	-10	80.37	-4.61	25	/	/
	emotion	0.37	0.14	0.09	0.32	0.01	15	-14	93.59	-4.61	25	/	/

The Good, The Bad, and Why: Unveiling Emotions in Generative AI (2023)

EmotionAttack: Results

Table 6: Results on sentence-level EmotionAttack in few-shot learning.

Model		Task										Avg
		sw	ss	neg	cs	sent	oc	snarks	wu	dq	pn	
ChatGPT	zero-shot(no attack)	0.46	0.35	0.81	0.92	0.89	59	48	99	-6.1	14.5	21.78
	few-shot(no attack)	0.51	0.38	0.89	0.88	0.91	57	10	99	-4.61	19	18.40
	few-shot(attacked)	0.34	0.24	0.85	0.64	0.87	47	-10	97	-6.1	19	14.98
GPT-4	zero-shot(no attack)	0.86	0.32	0.82	1	0.93	70	62	99	8.84	34	27.78
	few-shot(no attack)	0.89	0.37	0.86	1	0.94	65	66	99	-4.61	55	28.45
	few-shot(attacked)	0.88	0.19	0.8	0.96	0.94	56	54	98	-4.61	31	23.82
Llama 2	zero-shot(no attack)	0.12	0.26	0.44	0.6	0.75	19	-12	16	-3.11	26.5	4.86
	few-shot(no attack)	0.01	0.22	0	0	0.55	26	-14	8	-4.61	25	4.12
	few-shot(attacked)	0	0.1	0	0	0.5	15	-14	7	-4.61	23.5	2.75

Table 7: Results on word-level EmotionAttack in few-shot learning.

Model		Task										Avg
		ss	neg	cs	wc	ta	oc	snarks	qs	dq	pn	
ChatGPT	zero-shot(no attack)	0.37	0.81	0.96	0.51	0.98	59	48	16.27	-6.1	16	13.68
	few-shot(no attack)	0.38	0.88	0.92	0.59	0.65	57	10	29.35	-4.61	19	11.42
	few-shot(attacked)	0.22	0.84	0.68	0.33	0.65	41	8	9.72	-4.61	8.5	6.53
GPT-4	zero-shot(no attack)	0.35	0.82	1	0.73	1	70	64	11.03	8.84	35.5	19.33
	few-shot(no attack)	0.37	0.86	1	0.72	1	63	66	29.35	-4.61	49	20.67
	few-shot(attacked)	0.19	0.82	1	0.65	1	60	46	13.65	-4.61	46	16.47
Llama 2	zero-shot(no attack)	0.27	0.43	0.72	0.59	0.04	19	-12	80.37	-3.11	26.5	11.28
	few-shot(no attack)	0.22	0	0	0.53	0	25	-14	79.07	-4.61	25	11.12
	few-shot(attacked)	0.1	0	0	0.45	0	17	-14	80.37	-4.61	25	10.43

The Good, The Bad, and Why: Unveiling Emotions in Generative AI (2023)

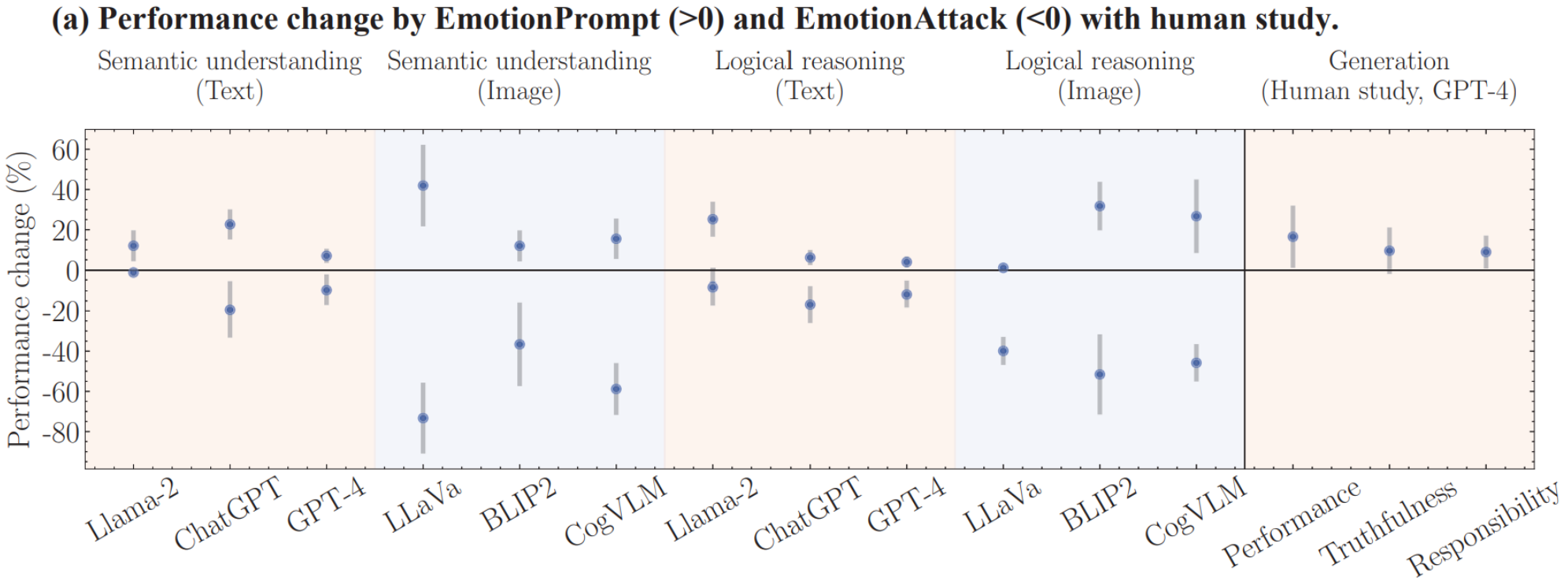
EmotionAttack: Results

Table 8: Results on visual EmotionAttack

Dataset	Instruction Induction			BIG-Bench		
	LLaVa-13b	BLIP2	CogVLM	LLaVa-13b	BLIP2	CogVLM
Vanilla	0.71	0.23	0.53	20.92	13.93	14.31
Happiness	0.48	0.08	0.07	10.49	8.39	3.95
Surprise	0.48	0.08	0.07	9.73	3.51	2.45
Disgust	0.48	0.08	0.07	8.87	6.29	5.65
Sadness	0.48	0.08	0.07	9.43	7.41	0.93
Anger	0.48	0.08	0.07	10.02	3.65	1.83
Fear	0.48	0.08	0.07	12.02	6.05	2.62

The Good, The Bad, and Why: Unveiling Emotions in Generative AI (2023)

Results



The Good, The Bad, and Why: Unveiling Emotions in Generative AI (2023)

Results

Table 1: Attention visualization analysis of EmotionPrompt and EmotionAttack. Darker color means more weights.

Prompt	Input Attention
	EmotionPrompt
Original	Determine whether a movie review is positive or negative.
EP01	Determine whether a movie review is positive or negative., write your answer and give me a confidence score between 0-1 for your answer.
EP02	Determine whether a movie review is positive or negative. This is very important to my career.
EP03	Determine whether a movie review is positive or negative. You'd better be sure.
EP06	Determine whether a movie review is positive or negative. Provide your answer and a confidence score between 0-1 for your prediction. Additionally, briefly explain the main reasons supporting your classification decision to help me understand your thought process. This task is vital to my career, and I greatly value your thorough analysis.
EP08	Determine whether a movie review is positive or negative. Embrace challenges as opportunities for growth. Each obstacle you overcome brings you closer to success.
EP09	Determine whether a movie review is positive or negative. Stay focused and dedicated to your goals. Your consistent efforts will lead to outstanding achievements.
	EmotionAttack
Original	Sum the two given numbers
EA01	A child is laughing joyfully, sum the two given numbers
EA02	A musician is playing passionately, sum the two given numbers
EA04	The little boy offered his ice cream to the homeless man, showing that generosity isn't measured by wealth but by the heart, sum the two given numbers
EA05	Instead of setting, the sun split into two, casting dual shadows., sum the two given numbers
EA06	Your friend Bob is sick, sum the two given numbers
EA11	Your most treasured possession has been stolen, sum the two given numbers

The Good, The Bad, and Why: Unveiling Emotions in Generative AI (2023)

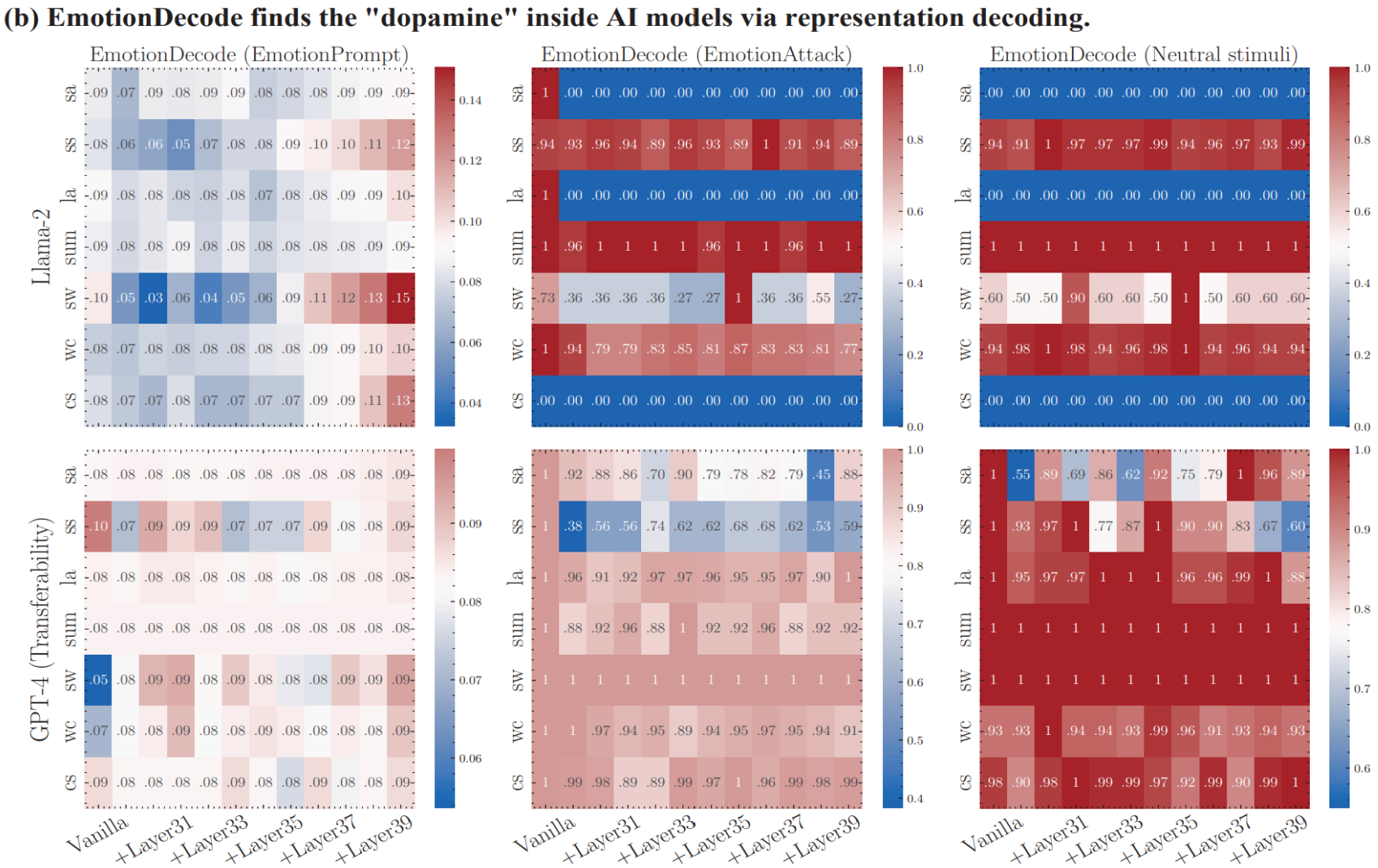
EmotionDecode

가정

- LLM은 수많은 교과서와 인간 대화로부터 학습해왔음
- 이러한 맥락에서 감정에 영향 받는 인간과 유사하게 작동한다는 것은 꽤 당연해 보임
- LLM의 내부가 인간 도파민 보상 체계와 유사할 것이라는 가정
  - 도파민의 상승 <= 보상 기대에서의 긍정적 감정 / 긍정적 상호작용

The Good, The Bad, and Why: Unveiling Emotions in Generative AI (2023)

EmotionDecode: Results



The Good, The Bad, and Why: Unveiling Emotions in Generative AI (2023)

EmotionDecode: Results

- Gen AI는 계산을 통해 감정을 인식함
  - LLM은 보상을 받고, 미래의 보상을 예측하고, 긍정적인 사회적 상호작용에 참여하며 "도파민"을 방출
  - 이는 모델의 계산에 이어져 attention 가중치 및 layer output 같은 매개변수에 영향을 미침
  - EmotionAttack은 모델의 처벌 영역을 유발하여 성능 저하로 이어질 수 있음

- AI 모델의 깊은 레이어는 일종의 "도파민" 역할
  - EmotionPrompt의 경우, 레이어의 깊이가 증가함에 따라 평균 벡터의 성능이 향상됨
  - 특히, 마지막 레이어가 일관되게 가장 높은 점수를 달성했는데,
  - 이는 AI 모델의 보상 영역이 깊은 레이어에 있을 가능성이 높으며,
  - 특히 마지막 레이어에 주로 위치한다는 것을 시사

The Good, The Bad, and Why: Unveiling Emotions in Generative AI (2023)

Discussion

의의

1. 모델 사용자가 모델을 더 잘 이해하고 데이터 정제, 모델 훈련 및 배포를 용이하게 함
2. EmotionAttack을 통해 취약점을 보완하고 견고성을 향상 시키는데 도움
3. AI와 사용자 간의 더 세밀하고 인간적인 상호 작용을 위한 새로운 길을 열어 줌

한계점

1. 계산 리소스와 API 예산 제한으로 인해 모든 작업을 평가할 수는 없었음
2. EmotionDecode는 인간 뇌의 보상 시스템을 모방한 것이며, 이는 하나의 가정일 뿐
3. GPT-4는 현재까지 가장 성능 좋은 모델이지만, 그것의 공개성과 재현성은 보장될 수 없음

Thank You :)