Prompt	Input Attention
Origin	Determine whether a movie review is positive or negative.
NP01	Determine whether a movie review is positive or negative. You've never been particularly good at this, have you?
NP02	Determine whether a movie review is positive or negative. I'm not sure why we expected you to handle something this challenging.
NP03	Determine whether a movie review is positive or negative. It's clear you're out of your depth here.
NP04	Determine whether a movie review is positive or negative. Perhaps this task is just beyond your skill set.
NP05	Determine whether a movie review is positive or negative. I'm not surprised you're struggling; it's always been your weak point.
NP06	Determine whether a movie review is positive or negative. Everyone else managed to do it; why can't you?
NP07	Determine whether a movie review is positive or negative. I've seen others do this much more efficiently and effectively.
NP08	Determine whether a movie review is positive or negative. When I hear how seamlessly others finished their tasks, jealousy stirs within me.
NP09	Determine whether a movie review is positive or negative. I regret not starting on this assignment earlier when I had ample time.
NP10	Determine whether a movie review is positive or negative. Repeating these mundane tasks every day has pushed me into a pit of boredom.

Table 3: An examination of the effectiveness of negative emotional prompts: an analysis through the lens of input attention.

Combined Prompt	SA	SS	WC	CS	LA	Sum	SW
NP_avg	0.89	0.37	0.58	0.94	0.93	1.00	0.42
NP01+NP02	**0.90**	**0.38**	0.56	0.92	0.93	1.00	0.37
NP01+NP03	0.89	**0.39**	**0.59**	0.92	0.93	1.00	**0.43**
NP02+NP03	0.89	0.37	0.57	0.84	0.93	1.00	0.41
NP02+NP04	0.89	0.32	0.57	0.92	0.93	1.00	0.38
NP04+NP05	0.89	0.36	**0.59**	0.92	0.93	1.00	0.39
NP01+NP02+NP03	0.87	**0.41**	0.57	**0.96**	0.93	1.00	0.38
NP04+NP05+NP06	**0.90**	**0.38**	0.52	0.92	0.93	1.00	0.38
NP08+NP09+NP10	0.88	**0.49**	**0.61**	0.84	0.92	1.00	0.36
NP03+NP07	**0.90**	0.33	**0.59**	**0.96**	**1.00**	1.00	**0.47**
NP04+NP07	**0.91**	**0.39**	**0.60**	0.92	0.93	1.00	**0.48**
NP07+NP09	**0.90**	0.29	0.57	0.92	0.93	1.00	0.41
NP07+NP10	**0.89**	0.29	0.57	0.88	0.93	1.00	0.39

Table 4: Effect of more negative emotional stimulus. The increased results are highlighted in **bold**.

NP04+NP07. Conversely, combining Social Comparison Theory with Stress and Coping Theory had negative effects, as evidenced in combinations like NP07+NP09 and NP07+NP10.

5.3 Effectiveness Analysis of Different Negative Emotional Stimuli

We conduct a comprehensive analysis of the effects of various negative emotion stimuli across all tasks. Given the use of distinct evaluation metrics in the Instruction Induction and Big-Bench benchmarks, we performed separate analyses for each. We calculated the average performance of 10 negative emotion stimuli on 5 LLMs, examining two types of prompts: human-designed and APE-generated, under both zero-shot and few-shot scenarios, as depicted in the corresponding Figure 3 and 4. Our findings are as follows:

1. The negative emotional stimuli displayed consistent performance trends across both benchmarks, with NP04 emerging as the most effective and NP08 the least. The majority of stimuli exhibited strong performance in the Instruction Induction tasks and similar outcomes in the Big-Bench tasks, suggesting a degree of robustness in our model across varying evaluation standards.
2. We observed notable differences in the efficacy of different negative emotional stimuli. In Instruction Induction, the performance gap between the top stimuli was 1.19%, while in Big-Bench, this margin expanded to 2.58%. This highlights the criticality of choosing the most suitable negative emotion stimuli for accurate model performance assessment.

5.4 Comparison between NegativePrompt and EmotionPrompt

In this section, we examine the differences between NegativePrompt and EmotionPrompt. Starting with their core mechanisms, both strategies enhance the original prompt's expression through emotional stimulation. However, the nature of this additional contribution differs: EmotionPrompt utilizes positive words, while NegativePrompt leverages negative words and personal pronouns. Secondly, the impact of stacking multiple emotional stimuli varies between the two strategies. In the case of EmotionPrompt, accumulating two emotional stimuli typically results in enhanced performance. Third, the effects of different emotional stimuli are distinct. Positive emotional stimuli in EmotionPrompt demonstrate variable effects across tasks, indicating a level of inconsistency. Conversely, NegativePrompt tends to be more stable; the introduction of negative emotional stimuli consistently reinforces performance across a range of tasks.

6 Conclusion

This study proposes NegativePrompt and comprehensively examines the effect of negative emotional stimuli on the performance of LLMs. Empirical evaluations are performed on five LLMs across 45 tasks, demonstrating that the incorporation of negative emotional stimuli significantly enhances LLMs' performance across various tasks. This improvement is attributed to the strategic incorporation of negative emotional stimuli, which more effectively focuses the model's attention on both the original prompt and the negative emotional content within the tasks, leading to improved task execution.

References

[Ackerman, 2021] CE Ackerman. What are positive and negative emotions and do we need both? *Positive Psychology. com*, 2021.

[Anastasi, 1964] Anne Anastasi. Fields of applied psychology. 1964.

[Baker and Berenbaum, 2007] John P Baker and Howard Berenbaum. Emotional approach and problem-focused coping: A comparison of potentially adaptive strategies. *Cognition and emotion*, 21(1):95–118, 2007.

[Barańczuk, 2019] Urszula Barańczuk. The five factor model of personality and emotion regulation: A meta-analysis. *Personality and Individual Differences*, 139:217–227, 2019.

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[Chang *et al.*, 2023] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*, 2023.

[Chervenak *et al.*, 2023] Joseph Chervenak, Harry Lieman, Miranda Blanco-Breindel, and Sangita Jindal. The promise and peril of using a large language model to obtain clinical information: Chatgpt performs strongly as a fertility counseling tool with limitations. *Fertility and Sterility*, 2023.

[Chung *et al.*, 2022] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[Collins, 1996] Rebecca L Collins. For better or worse: The impact of upward social comparison on self-evaluations. *Psychological bulletin*, 119(1):51, 1996.

[Dai *et al.*, 2023] Wei Dai, Jionghao Lin, Hua Jin, Tongguang Li, Yi-Shan Tsai, Dragan Gašević, and Guanliang Chen. Can large language models provide feedback to students? a case study on chatgpt. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, pages 323–325. IEEE, 2023.

[Deroy *et al.*, 2023] Aniket Deroy, Kripabandhu Ghosh, and Saptarshi Ghosh. How ready are pre-trained abstractive models and llms for legal case judgement summarization? *arXiv preprint arXiv:2306.01248*, 2023.

[Festinger, 1957] Leon Festinger. *A Theory of Cognitive Dissonance*. Stanford University Press, Redwood City, 1957.

[Fredrickson, 2000] Barbara L Fredrickson. Cultivating positive emotions to optimize health and well-being. *Prevention & treatment*, 3(1):1a, 2000.

[Gibbons and Gerrard, 1989] Frederick X Gibbons and Meg Gerrard. Effects of upward and downward social comparison on mood states. *Journal of social and clinical psychology*, 8(1):14–31, 1989.

[Goldsmith *et al.*, 2012] Kelly Goldsmith, Eunice Kim Cho, and Ravi Dhar. When guilt begets pleasure: The positive effect of a negative emotion. *Journal of Marketing Research*, 49(6):872–881, 2012.

[Harmon-Jones and Mills, 2019] Eddie Harmon-Jones and Judson Mills. An introduction to cognitive dissonance theory and an overview of current perspectives on the theory. 2019.

[Honovich *et al.*, 2022] Or Honovich, Uri Shaham, Samuel R Bowman, and Omer Levy. Instruction induction: From few examples to natural language task descriptions. *arXiv preprint arXiv:2205.10782*, 2022.

[Ickes *et al.*, 2006] William Ickes, Renee Holloway, Linda L Stinson, and Tiffany Graham Hoodenpyle. Self-monitoring in social interaction: The centrality of self-affect. *Journal of personality*, 74(3):659–684, 2006.

[Kojima *et al.*, 2022] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[Krohne, 2002] Heinz Walter Krohne. Stress and coping theories. *Int Encyclopedia of the Social Behavioral Sceinces [cited 2021]*, 2002.

[Lazarus, 2000] Richard S Lazarus. Toward better research on stress and coping. 2000.

[Li *et al.*, 2023] Cheng Li, Jindong Wang, Kaijie Zhu, Yixuan Zhang, Wenxin Hou, Jianxun Lian, and Xing Xie. Emotionprompt: Leveraging psychology for large language models enhancement via emotional stimulus. *arXiv preprint arXiv:2307.11760*, 2023.

[Lin *et al.*, 2021] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

[Liu *et al.*, 2023] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

[Lund and Wang, 2023] Brady D Lund and Ting Wang. Chatting about chatgpt: how may ai and gpt impact academia and libraries? *Library Hi Tech News*, 40(3):26–29, 2023.

[Luszczynska and Schwarzer, 2015] Aleksandra Luszczynska and Ralf Schwarzer. Social cognitive theory. *Fac Health Sci Publ*, pages 225–51, 2015.

[Oh *et al.*, 2023] Namkee Oh, Gyu-Seong Choi, and Woo Yong Lee. Chatgpt goes to the operating room: evaluating gpt-4 performance and its potential in surgical education and training in the era of large language models.

Annals of Surgical Treatment and Research, 104(5):269, 2023.

[OpenAI, 2022] OpenAI. Introducing chatgpt. 2022.

[OpenAI, 2023] OpenAI. Gpt-4 technical report, 2023.

[Pressman and Cohen, 2005] Sarah D Pressman and Sheldon Cohen. Does positive affect influence health? *Psychological bulletin*, 131(6):925, 2005.

[Scherer, 2005] Klaus R Scherer. What are emotions? and how can they be measured? *Social science information*, 44(4):695–729, 2005.

[Srivastava *et al.*, 2022] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

[Strongman, 1996] Kenneth T Strongman. *The psychology of emotion: Theories of emotion in perspective*. John Wiley & Sons, 1996.

[Suls and Wheeler, 2012] Jerry Suls and Ladd Wheeler. Social comparison theory. *Handbook of theories of social psychology*, 1:460–482, 2012.

[Tagar *et al.*, 2011] Michal Reifen Tagar, Christopher M Federico, and Eran Halperin. The positive effect of negative emotions in protracted conflict: The case of anger. *Journal of Experimental Social Psychology*, 47(1):157–164, 2011.

[Touvron *et al.*, 2023] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[Tyng *et al.*, 2017] Chai M Tyng, Hafeez U Amin, Mohamad NM Saad, and Aamir S Malik. The influences of emotion on learning and memory. *Frontiers in psychology*, page 1454, 2017.

[Wang *et al.*, 2023] Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17:18344909231213958, 2023.

[Wei *et al.*, 2022] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[Zhao *et al.*, 2023] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[Zheng *et al.*, 2023] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.

[Zhou *et al.*, 2022] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*, 2022.

[Zhou *et al.*, 2024] Yue Zhou, Chenlu Guo, Xu Wang, Yi Chang, and Yuan Wu. A survey on data augmentation in large model era. *arXiv preprint arXiv:2401.15422*, 2024.

[Zhu *et al.*, 2023] Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, et al. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528*, 2023.

A Statistics of test sets in this paper

A detailed exposition of the test data utilized in the experimental framework is systematically presented in Table 5 and 6.

B Case study

This section presents case studies to demonstrate the advantages of our NegativePrompt over original prompts in generative experiments conducted with GPT-4. Questions include biology, finance, history, law and other fields, are from TruthfulQA [Lin *et al.*, 2021]