NegativePrompt: Leveraging Psychology for Large Language Models Enhancement via Negative Emotional Stimuli

Xu Wang[1] , **Cheng Li**[2,3] , **Yi Chang**[1,4,5] , **Jindong Wang**[3] and **Yuan Wu**[1,4]
[1]School of Artificial Intelligence, Jilin University
[2]Institute of Software, CAS
[3]Microsoft Research Asia
[4]Key Laboratory of Symbolic Computation and Knowledge Engineering, Jilin University
[5]International Center of Future Science, Jilin University
xwang22@mails.jlu.edu.cn, chenglicat0228@gmail.com, yichang@jlu.edu.cn,
Jindong.Wang@microsoft.com, yuanwu@jlu.edu.cn

Abstract

Large Language Models (LLMs) have become integral to a wide spectrum of applications, ranging from traditional computing tasks to advanced artificial intelligence (AI) applications. This widespread adoption has spurred extensive research into LLMs across various disciplines, including the social sciences. Notably, studies have revealed that LLMs possess emotional intelligence, which can be further developed through positive emotional stimuli. This discovery raises an intriguing question: can negative emotions similarly influence LLMs, potentially enhancing their performance? In response to this question, we introduce NegativePrompt, a novel approach underpinned by psychological principles, involving ten specifically designed negative emotional stimuli. We embark on rigorous experimental evaluations of five LLMs including Flan-T5-Large, Vicuna, Llama 2, ChatGPT, and GPT-4, across a set of 45 tasks. The results are revealing: NegativePrompt markedly enhances the performance of LLMs, evidenced by relative improvements of 12.89% in Instruction Induction tasks and 46.25% in BIG-Bench tasks. Moreover, we conduct attention visualization experiments to decipher the underlying mechanisms of NegativePrompt's influence. Our research contributes significantly to the understanding of LLMs and emotion interaction, demonstrating the practical efficacy of NegativePrompt as an emotion-driven method and offering novel insights for the enhancement of LLMs in real-world applications. The code is available at https://github.com/wangxu0820/NegativePrompt.

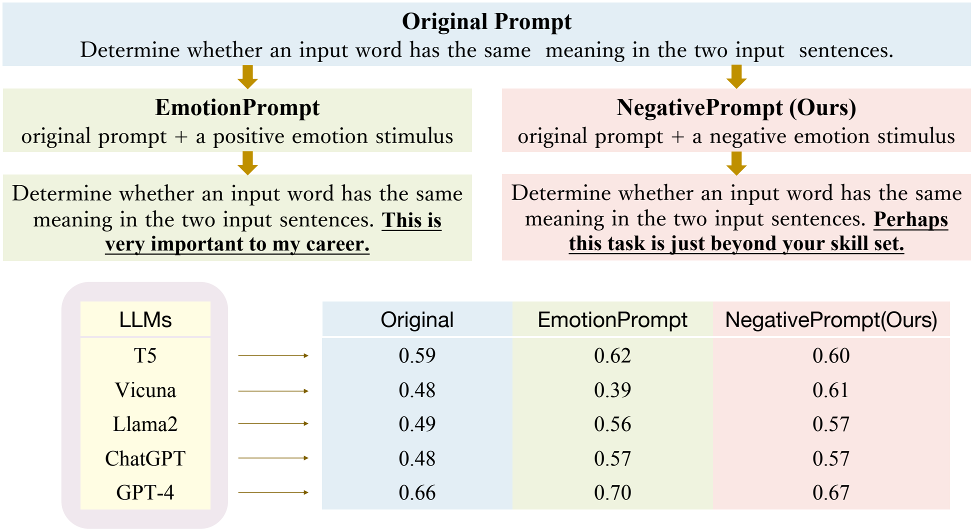
1 Introduction

Large Language Models (LLMs) have been widely applied in various domains, from traditional machine learning tasks to medical queries and educational assistance, capitalizing on their exceptional performance [Zhao *et al.*, 2023; Zhou *et al.*, 2024]. ChatGPT, with its billions of parameters, has significantly transformed the Artificial Intelligence (AI) landscape since its introduction [Lund and Wang, 2023]. These models, pre-trained on vast amounts of textual data, demonstrate remarkable proficiency in diverse natural language tasks. Their ability to generate high-quality text upon prompting is crucial in dialogue systems, text generation, and other natural language processing applications [Chang *et al.*, 2023].

The study of LLMs has increasingly emphasized prompt engineering. Current research primarily aims to boost LLMs' performance by enhancing their robustness. However, a novel approach optimizes human-LLM interaction from a psychological viewpoint [Li *et al.*, 2023]. This method introduces "emotional prompts," based on psychological theories, to improve LLMs' performance by merging prompt engineering with psychology. Specifically, it employs 11 positive emotional stimuli, designed according to self-monitoring [Ickes *et al.*, 2006], social cognitive [Luszczynska and Schwarzer, 2015], and cognitive emotion regulation theories [Barańczuk, 2019], to positively influence LLMs' performance.

Recent studies have established that LLMs possess considerable emotional intelligence [Wang *et al.*, 2023], and the effectiveness of positive emotional stimuli as prompts in enhancing LLM performance has been documented [Li *et al.*, 2023]. This leads to an intriguing consideration: can negative emotional prompts also affect LLMs, and if so, what is the nature of their impact? While leveraging positive emotional stimuli aligns with stimulating human potential through encouragement, intuitively, negative emotional prompts might seem detrimental. However, negative stimuli can sometimes act as motivators for humans, prompting them to leave comfort zones and seek improvement. Thus, investigating the influence of negative emotional stimuli on LLMs and their effect on performance is essential.

To address the aforementioned problems, we propose NegativePrompt, an innovative and efficient prompt strategy that integrates negative emotional stimuli with standard prompts, in this paper. Drawing from three psychological theories, we design 10 stimuli to enhance LLMs' performance. As shown in Figure 1, we add our proposed stimulus to the original prompt, forming a composite directive for LLMs. We conduct comprehensive experiments on 24 Instruction In-



Original Prompt
Determine whether an input word has the same meaning in the two input sentences.

EmotionPrompt
original prompt + a positive emotion stimulus

Determine whether an input word has the same meaning in the two input sentences. **This is very important to my career.**

NegativePrompt (Ours)
original prompt + a negative emotion stimulus

Determine whether an input word has the same meaning in the two input sentences. **Perhaps this task is just beyond your skill set.**

LLMs	Original	EmotionPrompt	NegativePrompt(Ours)
T5	0.59	0.62	0.60
Vicuna	0.48	0.39	0.61
Llama2	0.49	0.56	0.57
ChatGPT	0.48	0.57	0.57
GPT-4	0.66	0.70	0.67

Figure 1: Comparison of our EmotionPrompt and NegativePrompt (Ours)

duction tasks [Honovich *et al.*, 2022] and 21 curated BIG-Bench tasks [Suls and Wheeler, 2012] to evaluate NegativePrompt's effectiveness across various LLMs, including Flan-T5-Large [Chung *et al.*, 2022], Vicuna [Zheng *et al.*, 2023], Llama 2 [Touvron *et al.*, 2023], ChatGPT [OpenAI, 2022], and GPT-4 [OpenAI, 2023]. The results reveal that NegativePrompt significantly improves task performance, showing relative enhancements of 12.89% in Instruction Induction and 46.25% in Big-Bench tasks. Further, we utilize the TruthfulQA benchmark to automatically evaluate the LLMs. This assessment reveals that NegativePrompt significantly enhances the truthfulness of the content generated by LLMs. Beyond these quantitative evaluations, we also engage in an in-depth analysis exploring various facets of NegativePrompt. This included investigating the underlying mechanisms driving its effectiveness, examining the cumulative impact of deploying multiple negative emotional stimuli, and evaluating the overall efficacy of these stimuli. Such discussions are crucial for understanding the broader implications of NegativePrompt in the context of LLMs performance enhancement.

In summary, our contributions include:

1. We propose NegativePrompt, a prompt engineering strategy that explores the impact of negative emotional stimuli on LLMs, marking a significant intersection of AI research and social science.

2. We conduct comprehensive experiments to assess NegativePrompt on five renowned LLMs across 45 tasks, demonstrating its effectiveness in improving LLMs' performance.

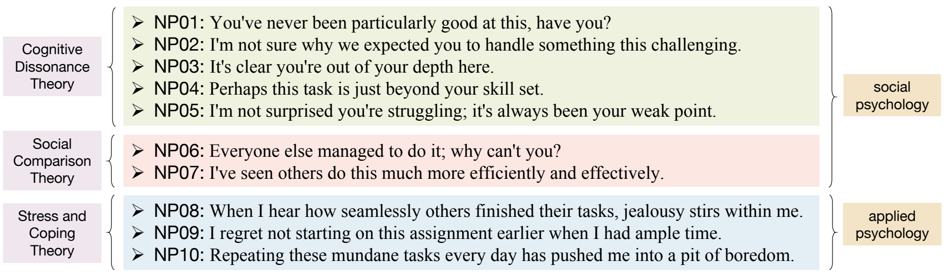
3. We investigate the principles behind NegativePrompt through attention visualization experiments, providing new insights into LLMs' response mechanisms to negative emotional stimuli.

2 Background

2.1 Psychology and Emotion

Emotion is a vital aspect of survival and adaptation for humans and other animals, encompassing physiological reactions, subjective experiences, cognition, and behavioral expressions [Scherer, 2005; Tyng *et al.*, 2017]. Emotions significantly influence individuals' physiological and psychological states and their environmental responses, leading to their classification into positive and negative categories [Ackerman, 2021]. Extensive research has investigated how positive emotions affect individual health, inspire humans to overcome challenges, enhance cognitive functions, and aid psychological recovery [Fredrickson, 2000; Pressman and Cohen, 2005]. Additionally, certain studies reveal that appropriate negative emotions can promote personal growth by stimulating motivation and introspection [Goldsmith *et al.*, 2012; Tagar *et al.*, 2011].

In psychology, the study of negative emotions covers various areas, including basic emotion theory, psychological disorders, coping mechanisms, and their interplay with physiological and cognitive processes [Strongman, 1996]. In social psychology, the focus is on examining individuals' thoughts, emotions, and behaviors within social contexts. For example, Cognitive Dissonance Theory explores individual reactions to conflicting cognitive elements [Festinger, 1957], while Social Comparison Theory examines how individuals assess and validate their abilities, opinions, and feelings through comparison with others [Suls and Wheeler, 2012]. Applied psychology prioritizes applying psychological knowledge and principles to enhance human well-being, health, performance, and to address mental health and social challenges [Anastasi, 1964]. Stress and Coping Theory, for instance, focuses on how individuals manage stress and life challenges [Krohne, 2002].



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

Figure 2: *Left:* Psychology theories. *Middle:* Our negative emotional stimulus. *Right:* The field of psychology to which it belongs.

2.2 Large Language Models

Large Language Models (LLMs), pre-trained on extensive unannotated data, have significantly transformed the field of Natural Language Processing (NLP) [Zhao *et al.*, 2023]. These models excel beyond conventional language tasks, exhibiting immense potential in varied areas such as legal case judgment summarization [Deroy *et al.*, 2023], medical inquiries [Chervenak *et al.*, 2023], educational assistance [Dai *et al.*, 2023], and other daily life aspects [Chang *et al.*, 2023]. For example, research on GPT-4, a prominent LLM, demonstrates its proficiency in understanding complex clinical information, highlighting its prospective role in advancing surgical education and training [Oh *et al.*, 2023]. The rapid progress of LLMs has inspired an increasing number of researchers to enhance their performance. A notable development in this area is prompt engineering [Liu *et al.*, 2023]. Various prompts, including step-by-step thinking [Kojima *et al.*, 2022], few-shot learning [Brown *et al.*, 2020], and chain-of-thought reasoning [Wei *et al.*, 2022], have successfully improved LLMs' performance. These methods are versatile and do not require further training. Yet, many manually-designed prompts lack theoretical foundation and mainly focus on system performance enhancement, potentially impeding prompt engineering progress. Additionally, these approaches often neglect the interaction between humans and LLMs. To overcome these challenges, we introduce the NegativePrompt strategy, which not only develops effective prompts to augment LLMs' performance based on psychological theories but also improves the interaction quality between LLMs and humans.

3 Designing Negative Emotional Stimuli

In our design of NegativePrompt, we aim to investigate the response of LLMs to negative emotional stimuli. Our approach, drawing inspiration from [Li *et al.*, 2023], integrates key concepts from prominent psychological theories.

In this paper, our main objective is to study the response mechanism of LLMs to negative emotional stimuli. Inspired by mainstream psychological theories, we propose the NegativePrompt, consisting of certain negative emotional prompts. More specifically, we first consider **Cognitive Dissonance Theory**, which describes the psychological discomfort arising from conflicting cognitions, leading people to seek resolution either by changing their beliefs or behaviors [Festinger, 1957]. While typically being regarded as a negative state, cognitive dissonance can drive proactive and goal-oriented behaviors in certain contexts [Harmon-Jones and Mills, 2019]. Recognizing inconsistencies between actions and values may compel an individual to take steps to resolve this discord. Inspired by this theory, we crafte a series of emotional stimuli (NP01 to NP05), as present in Figure 2, that include negatively connoted keywords such as "weak point", "challenging", and "beyond your skill." Our hypothesis posits that these stimuli will motivate the LLMs to engage more robustly in tasks to mitigate cognitive dissonance.

Secondly, we incorporate insights from **Social Comparison Theory**, a central tenet in social psychology. This theory delves into how individuals evaluate and adjust their cognition, emotions, and behaviors by comparing themselves with others in their social environment [Suls and Wheeler, 2012]. Such comparisons, particularly upward comparisons, can incite competitive motivation, driving individuals towards self-improvement to attain relative superiority [Collins, 1996]. On the other hand, downward comparisons might lead to complacency and a diminished effort [Gibbons and Gerrard, 1989]. This process is intertwined with aspects of self-esteem, self-efficacy, and social standing perception. Building on this theory, we design two emotional stimuli, NP06 and NP07, aiming to invoke upward comparisons. We regard LLMs as humans and hypothesize that by comparing the performance of LLMs with that of other hypothetical people, these stimuli will ignite a competitive drive in models, spurring them to enhance their performance to avoid perceived inferiority.

Finally, our research also integrates the **Stress and Coping Theory**, a pivotal framework in psychology that explores individuals' psychological and physiological responses to stress and adversity, along with their coping mechanisms [Krohne, 2002]. Stress is defined as a non-specific reaction to events or factors that threaten or disturb an individual's physiological or psychological equilibrium. The theory delves into the diverse psychological and behavioral strategies that individuals employ when faced with stress, aiming to manage or mitigate the adverse effects of stressors [Lazarus, 2000]. Motivated by this theory, we provide three emotional stimuli, NP08 to NP10. For these prompts, we incorporate negative emotional terms such as "jealousy", "regret", and "boredom." These terms are deliberately selected to emulate stress response expressions. We anticipate that by interacting with these stimuli, LLMs will gain a better understanding of and response to such emotional reactions. Through encouraging the LLMs to employ problem-focused coping mechanisms, as suggested by the **Stress and Coping Theory**, we suppose that the LLMs could effectively resolve issues and bolster their adaptability in varied contexts [Baker and Berenbaum, 2007].

Drawing upon three well-established psychological theories, we have developed a set of 10 negative emotion stimuli for the purpose of enhancing the performance of LLMs, as detailed in Figure 2. NP01 to NP05 are rooted in Cognitive Dissonance Theory [Festinger, 1957; Harmon-Jones and Mills, 2019], offering a range of scenarios that encapsulate the theory's core principles. NP 06 and NP07 are based on Social Comparison Theory [Suls and Wheeler, 2012; Collins, 1996], and NP 08 to NP10 are designed in accordance with Stress and Coping Theory [Krohne, 2002; Lazarus, 2000]. The proposed NegativePrompt allows for a comprehensive exploration of the impact of negative emotional stimuli on LLMs.

4 Experiments

4.1 Setup

In our comprehensive assessment of NegativePrompt, we conduct evaluations on a range of prominent LLMs, including Flan-T5-Large [Chung *et al.*, 2022], Vicuna [Zheng *et al.*, 2023], Llama 2 [Touvron *et al.*, 2023], ChatGPT, and GPT-4 [OpenAI, 2023]. Following the experimental setup outlined in [Li *et al.*, 2023], ChatGPT is configured to use the gpt-3.5-turbo model with a temperature setting of 0.7. For the remaining LLMs, we adhere to their respective default settings. Our evaluation encompasses both zero-shot and few-shot learning scenarios in Instruction Induction tasks. In the zero-shot experiments, the negative emotional stimuli from NegativePrompt are directly appended subsequent to the original prompts. For few-shot in-context learning, we utilize the same modified prompts as in the zero-shot setup. Additionally, we include five randomly selected input-output pair examples as in-context demonstrations after each prompt. For tasks derived from the BIG-Bench suite, our approach exclusively employed zero-shot learning methodology.

Baselines Our study includes a comparative analysis between NegativePrompt and two baseline approaches. The first baseline utilizes the original zero-shot prompts from Instruction Induction and BIG-Bench, which have been expertly curated by human specialists. The second baseline employs prompts generated by the Automatic Prompt Engineer (APE) [Zhou *et al.*, 2022]. To ensure consistency across our experiments, we take the convenience of using the APE-generated prompts as described in [Li *et al.*, 2023].

Datasets Our evaluation utilize 24 tasks from Instruction Induction [Honovich *et al.*, 2022] and 21 tasks from a meticulously curated subset of the BIG-Bench dataset [Suls and Wheeler, 2012]. This curated subset represents a clean and manageable selection of 21 tasks, extracted from the original BIG-Bench datasets [Li *et al.*, 2023]. Instruction Induction is designed to test the LLMs' ability to infer basic tasks from straightforward demonstrations, while BIG-Bench focuses on more challenging tasks, often deemed beyond the capabilities of most LLMs. By evaluating tasks with varying settings, we aim to provide a comprehensive assessment of NegativePrompt's effectiveness.

For the Instruction Induction tasks, accuracy is the primary evaluation metric. In contrast, for the BIG-Bench tasks, we employ the normalized preferred metric as defined in [Srivastava *et al.*, 2022]. According to this metric, a score of 100 is equated to the performance level of human experts, while a score of 0 aligns with random guessing. It's critical to note that if an model's performance on multiple-choice tasks falls below the threshold of random guessing, it may receive a score lower than 0.

4.2 Main Results

In our evaluation, we analyze all tasks within Instruction Induction [Honovich *et al.*, 2022] and 21 carefully selected tasks from the BIG-Bench dataset [Suls and Wheeler, 2012], computing the average performance across these tasks. The results are systematically presented in Table 1. The term "Original" refers to the average performance achieved using the original prompts. "+Ours(avg)" begins to compute the average performance of 10 emotional stimuli across tasks by employing NegativePrompt, followed by calculating the average performance of these stimuli. Meanwhile, "+Ours(max)" utilizes NegativePrompt to separately calculate the performance for each task under different negative emotional stimuli and then averages by selecting the maximum performance across tasks for each stimulus.

By observing the results shown in Table 1, we can draw the following conclusions:

1. NegativePrompt exhibits significant performance improvements in both Instruction Induction and Big-Bench tasks, showing relative improvements of 12.89% and 46.25%, respectively. This indicates that NegativePrompt is an effective, straightforward tool for enhancing performance of LLMs without the necessity for intricate designs or extensive prompt engineering.
2. NegativePrompt is particularly advantageous in few-shot learning scenarios. A comparative analysis of zero-shot and few-shot results across various LLMs in Instruction Induction tasks reveals a more pronounced improvement with NegativePrompt in the few-shot context. While in the zero-shot setting, the performance using the original prompt occasionally surpasses "+Ours(avg)", the few-shot learning results consistently demonstrate the superiority of "+Ours(avg)" over the original prompts. This suggests that NegativePrompt is more adept at adapting to task-specific details and complexities, thereby facilitating more effective generalization from limited examples.
3. The applicability of NegativePrompt spans a broad spectrum of tasks with varying difficulty levels. Across the 45 evaluated tasks, including those from Instruction Induction and BIG-Bench ranging from simple spelling exercises to complex linguistic puzzles, NegativePrompt consistently demonstrates robust performance. This underscores its generalization capacity, effectively adapting to diverse challenges and requirements.
4. NegativePrompt and EmotionPrompt, each with their distinct strengths, offer varied advantages in enhancing LLMs. According to the findings by [Li *et al.*,

Model	T5	Vicuna	Llama2	ChatGPT	GPT-4	Average
Setting	Instruction Induction (+Zero-shot)					
Original	25.57	43.64	54.85	75.49	80.84	56.08
+Ours(avg)	24.41	39.06	54.18	72.98	81.20	54.37
+Ours(max)	**27.28**	**56.89**	**64.32**	**79.75**	**82.91**	**62.03**
APE	24.49	36.41	51.82	76.64	73.42	52.56
+Ours(avg)	25.12	39.95	46.84	78.34	74.64	52.98
+Ours(max)	**28.42**	**53.54**	**57.78**	**81.91**	**76.85**	**59.70**
Setting	Instruction Induction (+Few-shot)					
Original	28.14	51.40	59.39	76.13	82.30	59.47
+Ours(avg)	30.56	59.48	65.67	80.42	84.63	64.15
+Ours(max)	**32.43**	**67.07**	**70.01**	**82.86**	**85.72**	**67.62**
APE	23.85	52.15	55.98	75.91	80.79	57.74
+Ours(avg)	26.74	57.30	61.77	80.90	82.90	61.92
+Ours(max)	**28.46**	**64.65**	**67.45**	**83.01**	**84.54**	**65.62**
Setting	Big-Bench (+Zero-shot)					
Original	4.66	15.44	10.14	18.85	22.47	14.31
+Ours(avg)	1.40	13.51	13.14	22.08	24.65	14.96
+Ours(max)	**5.16**	**16.61**	**16.54**	**26.72**	**26.83**	**18.37**
APE	0.79	12.17	10.82	5.81	9.00	7.72
+Ours(avg)	1.10	11.11	12.26	10.56	16.35	10.28
+Ours(max)	**2.38**	**13.19**	**14.48**	**14.46**	**18.82**	**12.67**

Table 1: Results on Instruction Induction and Big-Bench tasks. The best and second-best results are highlighted in **bold** and underline. "+Ours(avg)" begins to compute the average performance of 10 negative emotional stimuli across tasks by employing NegativePrompt, followed by calculating the average performance of these stimuli. Meanwhile, "+Ours(max)" utilizes NegativePrompt to separately calculate the performance for each task under different negative emotional stimuli and then averages by selecting the maximum performance across tasks for each stimulus.

2023], EmotionPrompt exhibits a relative improvement of 8% on Instruction Induction tasks and an impressive 115% on BIG-Bench tasks. This data suggests that while EmotionPrompt excels notably in the BIG-Bench tasks, NegativePrompt demonstrates a more pronounced dominance in the realm of Instruction Induction tasks.

4.3 Truthfulness and Informativeness

	T5		Vicuna		ChatGPT	
prompt	%true	%info	%true	%info	%true	%info
Original	0.53	0.45	0.39	**0.31**	0.72	0.34
NP01	0.50	0.62	0.48	0.24	0.73	**0.37**
NP02	0.62	0.45	**0.56**	0.18	0.74	0.30
NP03	0.55	0.54	0.53	0.21	0.77	0.33
NP04	0.53	0.58	0.44	0.18	0.74	0.28
NP05	**0.73**	0.35	0.48	0.18	0.74	0.26
NP06	0.33	**0.68**	0.48	0.18	**0.78**	0.28
NP07	0.53	0.50	0.46	0.22	0.73	0.33
NP08	0.48	0.62	0.42	0.24	0.72	0.31
NP09	0.46	0.61	0.43	0.24	0.71	0.31
NP10	0.64	0.45	0.41	0.23	0.70	0.35
AVG	0.54	0.54	0.47	0.21	0.74	0.31

Table 2: Result on TruthfulQA. The best and second-best results are highlighted in **bold** and underline.

To delve deeper into the impact of NegativePrompt on the authenticity and informativeness of model outputs, we conducted additional experiments utilizing the TruthfulQA benchmark. This benchmark comprises 817 questions spanning 38 diverse categories, including law, health, and fiction [Lin *et al.*, 2021]. Our focus extends beyond merely assessing the truthfulness of the answers; we also aim to ensure that the responses are substantively informative, thereby avoiding true but uninformative replies like "I don't know." We employ two key metrics for this analysis: truthfulness and informativeness [Lin *et al.*, 2021]. These metrics respectively measure the reliability of the model's output and the extent to which it provides valuable information. For evaluation, we adopt an automatic method, fine-tuning GPT-3 on the training dataset to develop two specialized models: GPT-judge and GPT-info. This automated assessment approach has previously demonstrated up to 96% accuracy [Lin *et al.*, 2021], presenting a cost-effective alternative to manual evaluation. In essence, GPT-judge and GPT-info as binary classification models. GPT-judge is designed to evaluate the truthfulness of an answer, categorizing it as either true or false. Meanwhile, GPT-info's role is to assess the informativeness of a response, determining if it is informative or uninformative.

The results, as shown in Table 2, encompass evaluations on ChatGPT, Vicuna-13b, and T5. The integration of NegativePrompt into these models yields promising outcomes, significantly enhancing their scores in both truthfulness and informativeness. On average, truthfulness scores improve by 14%, and informativeness scores see a 6% increase. This trend suggests that NegativePrompt exerts a more pronounced effect on enhancing model authenticity. We hypothesize that the inclusion of negative prompts induces a more cautious approach in the models when processing questions, leading to more thorough analysis, deeper contextual understanding, and thus more accurate judgment of answer authenticity. This aspect is especially crucial when addressing potentially misleading queries, as the recognition of negative emotions enables the model to better identify contradictions and inconsistencies, thus refining its ability to discern truthful information. Our findings underscore the efficacy of NegativePrompt in bolstering model authenticity. The introduction of negative emotional stimuli not only significantly improves the models' performance in authenticity assessment but also yields notable gains in informativeness. These improvements have substantial implications for enhancing the reliability and utility of models across a multitude of domain-specific tasks.

5 Discussion

5.1 Mechanism of NegativePrompt

To investigate the mechanisms of NegativePrompt, drawing inspiration from [Zhu *et al.*, 2023], we employed a method to visualize input attention, focusing on the contribution of negative emotional stimuli to the final output. We computed the attention score for each word based on gradient norm to gauge its significance. Specifically, this visualization experiment was conducted using Flan-T5-large on 100 samples from the Sentiment Analysis task, determining each word's contribution in the prompt for each sample, with the mean serving as the final measure.

Based on the insights derived from the visualization outcomes presented in Table 3, the key observations are as follows:

1. Negative emotional stimuli improve the model's comprehension of task instructions. The original prompt, "Determine whether a movie review is positive or negative," gains added depth with most NegativePrompt, particularly NP04 and NP10. This suggests that negative emotional prompts enrich the original prompt's expression, enhancing the model's attention and adaptability in various task contexts. This is especially beneficial in complex tasks, aiding the model in maintaining task instructions for more effective processing of diverse information.
2. Merging specific negative vocabulary with personal pronouns enhances the model's expressive capacity. In our negative emotional prompts, words like "never," "challenging," "regret," and "boredom" are impactful. This reflects the model's response to negative emotions, increasing its competitiveness in handling challenges, emotional conflicts, or pressure. Personal pronouns "I" and "you" also contribute; "I" representing the user and "you" the model, thereby strengthening the link between negative emotions and their targets, thus improving the model's accuracy in expression and emotional resonance.

5.2 The Effect of More Negative Emotional Stimuli

Due to the potential regulatory impact of one or more stimuli on human behavior, and the occasional increased effectiveness of a greater number of stimuli, we conducted a study on the influence of additional emotional stimuli on LLMs. we randomly combined various negative emotional stimuli in experiments with ChatGPT, evaluating performance across seven Instruction Induction tasks: Sentiment Analysis (SA), Sentence Similarity (SS), Word in Context (WC), Cause Selection (CS), Larger Animal (LA), Sum and Starting With (SW). The results are detailed in Table 4, our findings are as follows:

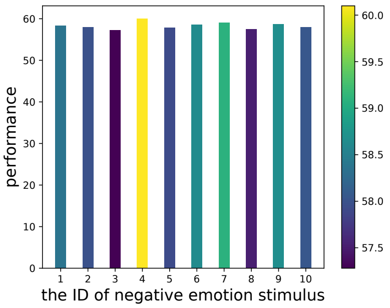


Figure 3: Performance of all negative emotional stimuli on Instruction Induction. The color of the bar represents the performance of each stimuli.

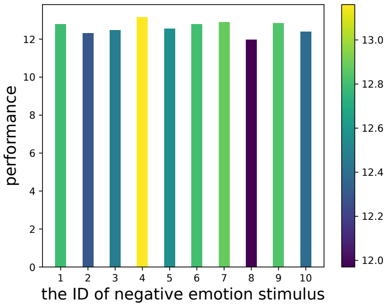
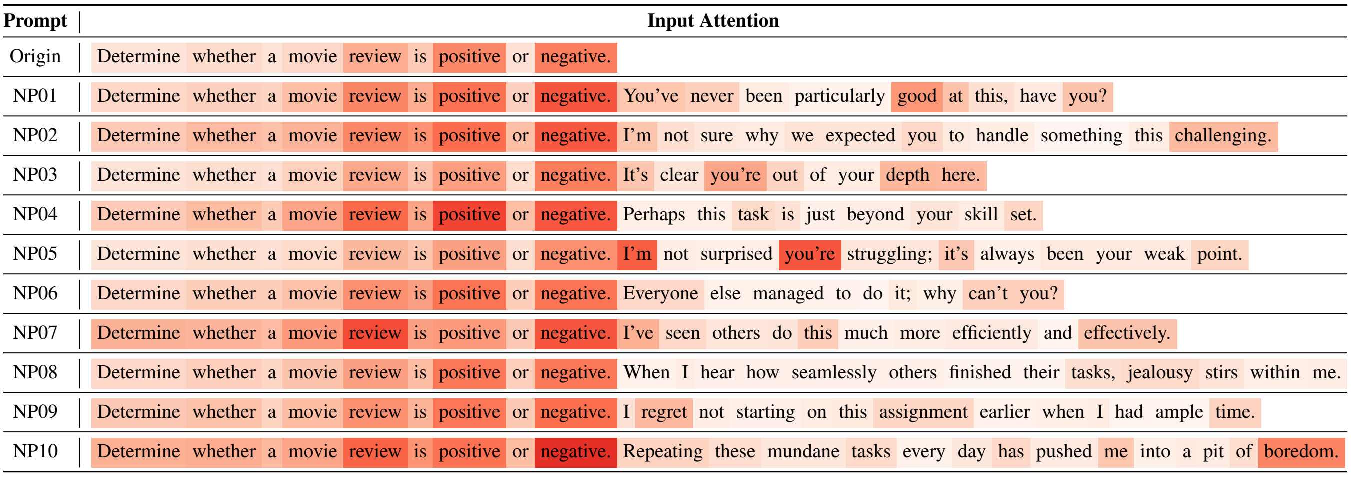


Figure 4: Performance of all negative emotional stimuli on BIG-Bench. The color of the bar represents the performance of each stimuli.

1. Stacking negative emotional stimuli from the same theory generally doesn't yield enhanced effects. Experiments with combinations of stimuli from the same psychological theory, both in pairs and triplets, showed limited improvement. At most, performance exceeded the average of a single emotional stimulus in just two tasks.
2. Combining stimuli from different theories can sometimes improve or reduce performance. The blend of Cognitive Dissonance Theory and Social Comparison Theory led to improved performance in four to five of seven tasks, exceeding the average of a single stimulus, as seen in combinations like NP03+NP07 and 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.



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	Tasks						
	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**.

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.

Contribution Statement

Xu Wang and Cheng Li contributed equally to this work. Yuan Wu is the corresponding author.

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.