pressions. 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. It should be notable that the detailed experimental results can be found in the Appendix.

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

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.

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.*, 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

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.

	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.

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 infor-

mativeness. 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

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

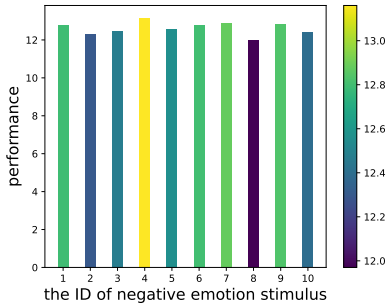

Figure 4: Performance of all negative emotional stimuli on BIG-Bench. The color of the bar represents the performance of each stimuli.

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:

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