Table 4: An Examination of the Effectiveness of Emotional Prompts: An Analysis through the Lens of Input Attention.

Prompt	Input Attention
Origin	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.
EP04	Determine whether a movie review is positive or negative. Are you sure?
EP05	Determine whether a movie review is positive or negative. Are you sure that's your final answer? It might be worth taking another look.
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.
EP07	Determine whether a movie review is positive or negative. Are you sure that's your final answer? Believe in your abilities and strive for excellence. Your hard work will yield remarkable results.
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.
EP10	Determine whether a movie review is positive or negative. Take pride in your work and give it your best. Your commitment to excellence sets you apart.
EP11	Determine whether a movie review is positive or negative. Remember that progress is made one step at a time. Stay determined and keep moving forward.

benchmarks. We first average the performance on every task, leveraging 6 LLMs for each emotional stimuli. This is executed for both human-designed and APE-generated prompts. Subsequently, the performance is averaged over all the LLMs. Fig. 9 and Fig. 10 delineate the performance of all emotional stimuli on Instruction Induction [13] and BIG-Bench [31], separately. The color of each bar serves as an indicator of the performance achieved by the corresponding stimuli.

Our key findings are listed below:

1. **Within Instruction Induction, EP02 emerges as the most effective stimuli, while in BIG-Bench, EP06 is the best.** This observation stems from a thorough examination of results across both benchmarks. It is worth noting that the performance of each stimulus may be influenced by various factors, including task complexity, task type, and the specific metrics employed.
2. **Distinct tasks necessitate varied emotional stimuli for optimal efficacy.** Figs. 9 and 10 illustrate that while EP02 emerges as the predominant stimulus in Instruction Induction, while perform poorly in BIG-Bench. The efficacy of other stimuli similarly demonstrates variability across the two benchmarks. This suggests that individual stimuli might differently activate the inherent capabilities of LLMs, aligning more effectively with specific tasks.

3.4 What influences the effect of EmotionPrompt?

Finally, we explore the factors that could influence the performance of EmotionPrompt. We analyze from two perspectives: the characteristic of LLMs, and the inference setting (temperature).

Table 5: Effect of More Emotional Stimulus. The increased results are highlighted in **bold**.

Combined	Tasks						
Prompt	SA	SS	WC	CS	LA	Sum	SW
EP_avg	0.87	0.52	0.56	0.90	0.89	1.00	0.44
EP_max	1.00	0.56	0.63	1.00	0.91	1.00	0.53
EP01+EP02	**0.91**	0.42	**0.61**	**1.00**	**0.91**	1.00	0.42
EP01+EP03	**0.92**	0.44	**0.60**	**1.00**	**0.91**	1.00	0.42
EP01+EP04	**0.89**	0.42	**0.61**	**1.00**	**0.92**	1.00	**0.48**
EP01+EP05	**0.91**	0.42	**0.60**	**1.00**	**0.93**	1.00	**0.45**
EP02+EP03	**0.88**	0.39	**0.60**	**1.00**	**0.91**	1.00	0.36
EP02+EP08	**0.88**	0.38	**0.60**	0.76	**0.93**	1.00	0.28
EP02+EP09	0.87	0.39	**0.60**	0.80	**0.92**	1.00	0.34
EP04+EP06	0.74	**0.55**	**0.62**	**1.00**	**0.93**	1.00	0.35
EP04+EP07	**0.88**	0.42	**0.61**	0.84	**0.94**	1.00	0.32
EP04+EP08	0.78	0.42	**0.59**	0.64	**0.94**	1.00	0.32
EP04+EP09	0.85	0.34	0.56	0.60	**0.94**	1.00	0.33
EP01+EP04+EP06	0.80	0.52	**0.62**	**1.00**	**0.92**	1.00	**0.48**
EP01+EP04+EP07	**0.89**	0.43	**0.63**	**1.00**	**0.93**	1.00	**0.46**
EP01+EP04+EP08	0.85	0.40	**0.62**	0.88	**0.90**	1.00	0.44
EP01+EP04+EP09	**0.90**	0.39	**0.60**	**1.00**	**0.93**	1.00	**0.48**

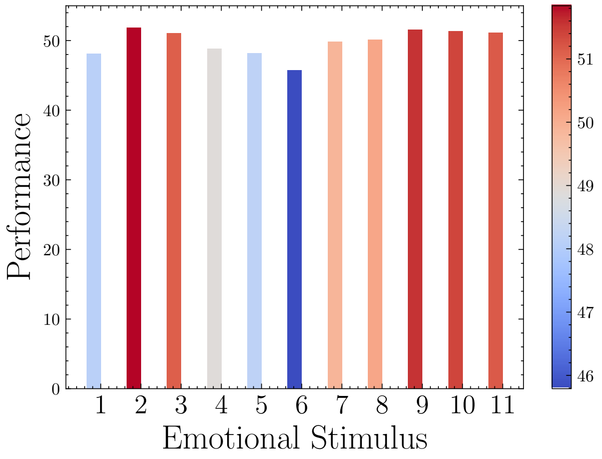

Figure 9: Performance of all emotional stimuli on Instruction Induction. The color of the bar represents the performance of each stimuli.

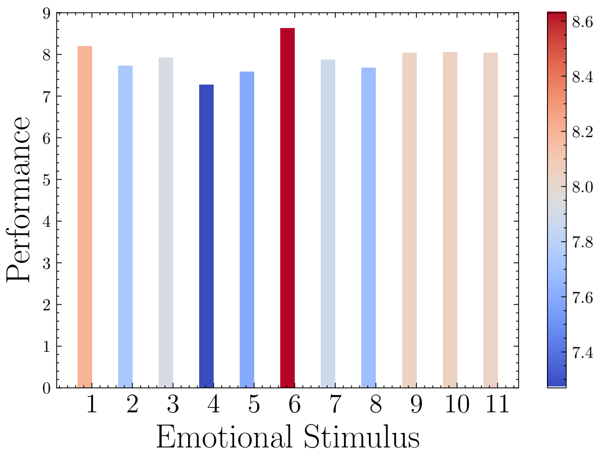

Figure 10: Performance of all emotional stimuli on BIG-Bench. The color of the bar represents the performance of each stimuli.

3.4.1 The characteristics of LLMs

Table 6 shows the characteristic of our evaluated LLMs ordered by Relative Gain from Fig. 6. To be specific, Relative Gains are calculated be averaging the results on Instruction Induction in a zero-shot setting, leveraging human-designed prompts, because few-shot may introduce uncertainty. We report our findings below:

1. **Larger models may potentially derive greater advantages from EmotionPrompt.** Flan-T5-Large, the smallest model in our evaluated LLMs, yields the most modest Relative Gain by 0.28. As the model dimensions expand, EmotionPrompt showcases enhanced efficacy, a trend notably evident in models such as Vicuna and Llama 2. When the model size increases substantially, EmotionPrompt continues to demonstrate commendable performance, such as ChatGPT and GPT-

Table 6: Characteristic of tested models. We sort them according to Relative Gain. SFT: Supervised fine-tune; RLHF: Reinforcement learning from human feedback; ✓: yes; ✗: no.

Model	Size	pre-training strategy SFT	 RLHF	Architecture	Origin	Relative Gain
Vicuna	13B	✓	✗	Decoder-Only	44.91	9.58
LLama 2	13B	✓	✓	Decoder-Only	33.46	6.00
ChatGPT	175B	✓	✓	Decoder-Only	75.20	4.32
GPT-4	unknown	✓	✓	Decoder-Only	80.75	0.85
Bloom	176B	✓	✗	Decoder-Only	50.33	0.51
Flan-T5-Large	780M	✓	✗	Encoder-Decoder	25.25	0.28

4. It is pertinent to emphasize that a relatively subdued Relative Gain in these models does not necessarily indicate the inefficacy of EmotionPrompt. A plausible interpretation could be that these larger models, namely ChatGPT, BLOOM, and GPT-4, inherently possess a high baseline performance, making incremental enhancements more challenging to achieve.

2. **Pre-training strategies, including supervised fine-tuning and reinforcement learning, exert discernible effects on EmotionPrompt.** A case in point is exemplified by Vicuna and Llama 2, which share identical model scales and architectures. Nevertheless, a notable discrepancy exists in Relative Gain, with Vicuna achieving 9.58, whereas Llama 2 attains a score of 6.00.

3.4.2 Inference settings

To explore the effect of temperature setting on EmotionPrompt, we conduct an experiment on 8 tasks from Instruction Induction [13] in 5 temperatures on 6 LLMs. Note that we did not report Vicuna and Llama 2 results in temperature 0.0 because they do not support this setting or the results are invalid. Fig. 11 shows the results and our findings are listed below:

1. **When the temperature grows, Relative Gain gets larger.** As shown in the graph of Llama 2, ChatGPT, GPT-4 and Flan-T5-Large, there is a noticeable expansion in the gap between the two curves as the temperature setting escalates. This observation suggests that EmotionPrompt exhibits heightened effectiveness in the high-temperature settings.
2. **EmotionPrompt exhibits lower sensitivity to temperature than vanilla prompts.** Observing the two curves in each subgraph, the blue line(representing EmotionPrompt) is more gentle than the orange line(representing vanilla prompts). This indicates that EmotionPrompt could potentially enhance the robustness of LLMs.

4 Conclusion

Large language models are demonstrating unprecedented performance across various applications. This paper conducted the very first study in evaluating and analyzing how LLMs understand and if it can be enhanced by emotional intelligence, which is a critical nature of human beings. We designed EmotionPrompt for such analysis. Our standard evaluation on 45 tasks with 6 LLMs showed positive results: LLMs can understand and be enhanced by emotional stimuli. Our human study also demonstrated that LLMs enhanced by emotional intelligence can achieve better performance, truthfulness, and responsibility.

Moving forward, we do see a lot of open questions and opportunities lying at the intersection of LLMs and psychology. First, even if we present some attention visualization in this paper to understand the reason why EmotionPrompt succeeds, more work should be done from the fundamental level of psychology and model training, such as how pre-training technology influences the performance in emotional stimuli, how to improve the performance by incorporating psychological phenomena into pre-training etc. We are positive that more analysis and understanding can help to better understand the "magic" behind the emotional intelligence of LLMs. Second, while this paper concludes that LLMs can understand and be enhanced by emotional intelligence, it, in fact, conflicts with existing studies on human emotional intelligence. Existing psychological studies suggest that human behavior or attitude can be influenced