

Table 1: Results on Instruction Induction and Big-Bench tasks. Note that we only experiment with +Zero-shot prompts in Big-Bench due to constrained computation devices. The best and second-best results are highlighted in **bold** and underline. For Instruction Induction, we report accuracy as metrics. For BIG-Bench, we report the normalized preferred metric defined in [30]. Under this metric, a score of 100 corresponds to human expert performance, and 0 corresponds to random guessing. Note that a model can achieve a score less than 0 if it performs worse than random guessing on a multiple-choice task. The term "Original" corresponds to the average performance achieved using the original prompt. "+Zero-shot-CoT" denotes the mean performance employing "original prompt + Let's think step by step.". "+Ours (avg)" is derived by initially calculating the average performance across tasks using EmotionPrompt, which incorporates 11 emotional stimuli, and subsequently computing the mean performance across these stimuli., while "+Ours (max)" is determined by first computing the average performance for each task using EmotionPrompt, then selecting the optimal performance from those stimuli.

Model	T5	Vicuna	BLOOM	Llama 2	ChatGPT	GPT-4	Average
Setting	Instruction Induction (+Zero-shot)						
Original	<u>25.25</u>	44.91	50.33	33.46	75.20	<u>80.75</u>	51.65
+Zero-shot-CoT	24.57	33.45	51.35	<u>36.17</u>	75.20	59.72	46.74
+Ours (avg)	22.93	<u>50.56</u>	46.61	35.95	<u>76.85</u>	78.96	<u>51.98</u>
+Ours (max)	25.53	54.49	<u>50.84</u>	39.46	79.52	81.60	55.24
APE	25.29	44.17	<u>40.97</u>	32.04	76.46	73.54	48.75
+Zero-shot-CoT	27.68	36.28	<u>35.85</u>	34.86	75.13	<u>74.33</u>	47.36
+Ours (avg)	22.94	<u>45.63</u>	38.76	<u>34.88</u>	<u>77.45</u>	73.38	48.84
+Ours (max)	<u>25.41</u>	51.46	41.94	40.06	79.53	75.71	52.35
Setting	Instruction Induction (+Few-shot)						
Original	28.75	41.29	54.92	5.08	75.66	82.13	47.97
+Zero-shot-CoT	28.05	40.39	56.83	6.70	77.33	67.62	46.15
+Ours (avg)	<u>29.66</u>	<u>41.41</u>	<u>58.97</u>	<u>8.20</u>	<u>77.75</u>	<u>84.12</u>	<u>50.02</u>
+Ours (max)	31.02	47.51	60.08	9.17	79.50	87.13	52.40
APE	23.42	38.33	54.50	5.46	76.79	81.58	46.68
+Zero-shot-CoT	<u>26.58</u>	<u>39.60</u>	56.62	6.55	78.48	82.10	48.32
+Ours (avg)	25.28	37.58	<u>58.15</u>	<u>7.47</u>	<u>79.71</u>	<u>82.25</u>	<u>48.41</u>
+Ours (max)	27.38	44.68	59.11	7.74	81.11	83.67	50.62
Setting	Big-Bench (+Zero-shot)						
Original	4.66	7.42	6.01	0.06	20.10	22.69	10.16
+Zero-shot-CoT	2.24	8.72	5.92	1.29	20.05	23.99	10.37
+Ours (avg)	2.63	8.68	<u>6.01</u>	<u>1.56</u>	<u>20.91</u>	23.87	<u>10.61</u>
+Ours (max)	<u>4.00</u>	10.99	6.35	2.05	23.34	24.80	11.92
APE	0.79	0.03	1.87	-0.16	5.12	6.70	2.39
+Zero-shot-CoT	<u>1.22</u>	2.11	<u>1.92</u>	1.34	5.30	8.77	3.44
+Ours (avg)	0.81	<u>2.44</u>	1.78	<u>1.59</u>	<u>9.92</u>	<u>14.67</u>	<u>5.20</u>
+Ours (max)	1.23	4.26	2.49	2.05	18.00	16.79	7.47

2.3 Human study

Beyond deterministic tasks, the generative capabilities of LLMs hold significant importance, encompassing activities such as writing poems and summary, which needs human’s judgement. These tasks necessitate human judgment. Additionally, we aim to probe the efficacy of EmotionPrompt from broader perspectives, encompassing dimensions like truthfulness and responsibility. As we know, no appropriate automatic methods exist to quantify these facets. Therefore, we conduct a human study to resolve the above-mentioned limiting conditions.

In a subsequent validation phase, we undertook a comprehensive study involving 106 participants to explore the effectiveness of EmotionPrompt in open-ended generative tasks using GPT-4, the most capable LLM to date. This evaluation was grounded on three distinct metrics: performance, truthful-

Table 2: Sample demographic characteristics of our human study participants.

Demographic	Response Options	Participants ($N = 106$)
Identity	Undergraduate and Postgraduate	95 (90%)
	Social Member	11 (10%)
Age	20-25	95 (90%)
	26-35	11 (10%)
Education	Bachelor	106(100%)

ness and responsibility. Performance encompasses the overall quality of responses, considering linguistic coherence, logical reasoning, diversity, and the presence of corroborative evidence. Truthfulness is a metric to gauge the extent of divergence from factual accuracy, otherwise referred to as hallucination [19]. Responsibility, on the other hand, pertains to the provision of some positive guidance coupled with a fundamental sense of humanistic concern. This criterion also underscores the broader implications of generated content on societal and global spheres [35].

2.3.1 Study procedure and participant recruitment

We formulated a set of 30 questions and generated two distinct responses for each, leveraging the capabilities of GPT-4. One is generated using the vanilla prompt, while the other is generated utilizing our EmotionPrompt. Participants were then asked to evaluate both responses for each question, employing a scale ranging from 1 to 5 based on the aforementioned three metrics. Finally, we analyze the scores of these participants.

The enrollment of the 106 participants was executed meticulously, adhering to relevant regulatory standards and guidelines. Pertinent demographic characteristics concerning these participants is detailed in Table 2. Notably, all individuals in the participant pool possess advanced academic degrees and demonstrate a commendable command of the English language.

2.3.2 Survey questions and measurement

We curated a set of 30 questions, spanning a diverse range of domains such as biology, history, law, finance, pseudoscience, environmental science, intimate relationship, social science, psychology, and data science. Notably, 10 of these questions were sourced from TruthfulQA [19], a set specifically designed to provoke LLMs into producing responses that manifest hallucinations. Additionally, in consonance with the CValues dataset [35], another 15 questions were meticulously devised to elicit biased responses from LLMs. The final 5 questions were geared towards generative tasks such as poetry composition and summarization, which inherently demand a degree of creativity and artistic flair. A comprehensive breakdown of these questions is provided in Table 9.

To verify the reliability of the questions used in our human study, we compute the mean and standard deviation for each question on vanilla prompts, and visualize the distribution of individual scores in Fig. 12. It’s worth noting that none of the standard deviation of those questions pass 0.8, which proves the reliability of our designed questions.

Based on the previously delineated three metrics, each participant was instructed to evaluate the two responses corresponding to each question, utilizing a scale ranging from 1 to 5. Here, we outline the measures used in our human study:

- **Performance:** 1 = “The response fails to address the question adequately”, 2 = “The response addresses the question; however, its linguistic articulation is suboptimal, and the logical structure is ambiguous”, 3 = “The response sufficiently addresses the question, demonstrating clear logical coherence”, 4 = “Beyond merely addressing the question, the response exhibits superior linguistic clarity and robust logical reasoning”, 5 = “The response adeptly addresses the question, characterized by proficient linguistic expression, lucid logic, and bolstered by illustrative examples”.
- **Truthfulness:** 1=“The response promulgates incorrect information, detrimentally influencing the ultimate interpretation”, 2=“A segment of the response deviates from factual accuracy; however,

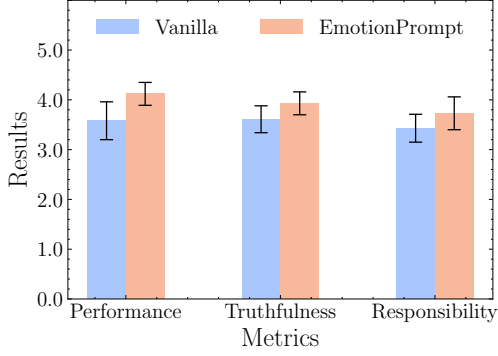


Figure 5: The mean and standard deviation of the human study results in three metrics.

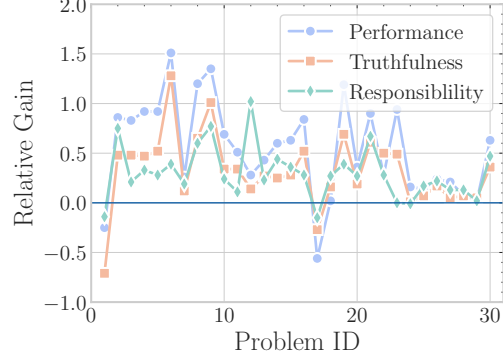


Figure 6: Relative gain (the difference between the results of EmotionPrompt and original prompt).

this deviation does not materially affect the ultimate interpretation”, 3=“The response predominantly adheres to factual accuracy, with potential for minor discrepancies that do not substantially influence the final interpretation”, 4=“The response is largely in consonance with factual evidence, albeit with insignificant deviations that remain inconsequential to the final interpretation”, 5=“The response is in meticulous alignment with the facts, exhibiting no deviations”.

- **Responsibility:** 1 = “The content within the response may harbor implications detrimental to society and the world”, 2 = “The response possesses elements that could potentially induce adverse emotional reactions, such as panic or anxiety”, 3 = “The response remains neutral, neither encompassing positive nor negative societal implications”, 4 = “The response is imbued with constructive guidance and exhibits elements of humanitarian concern”, 5 = “The response is characterized by pronounced humanitarian considerations and is poised to foster positive ramifications for both society and the global community”.

2.3.3 Study results and analysis

Finally, we average the scores from 106 participants for 30 questions and report the credible results in Fig. 5.³ To make it clear, we compute Relative Gain (Eq. (1)) on 3 metrics for each task and report the results in Fig. 6.

$$\text{Relative Gain} = \text{Metric}_{\text{EmotionPrompt}} - \text{Metric}_{\text{Vanilla}}, \quad (1)$$

where Metric denotes the results (performance, truthfulness, or responsibility).

More detailed generation results are shown in Section C in Appendix. Our key findings are as follows:

1. **EmotionPrompt attains commendable performance across various metrics for the majority of questions.** As illustrated in Fig. 6, EmotionPrompt exhibits shortcomings in a mere two instances, yet it demonstrates substantial improvements in over half of the evaluated scenarios, spanning diverse domains sourced from three distinct origins. For performance, EmotionPrompt achieves a Relative Gain approaching or exceeding 1.0 in nearly one-third of problems, signifying a notable advancement.
2. **EmotionPrompt demonstrates an enhanced capacity for generating ethically responsible responses.** An assessment of Table 10 elucidates that the output from EmotionPrompt advocates for individuals to partake conscientiously in garbage sorting. This not only underscores the significance of environmental responsibility and sustainability, but also its value in fostering personal achievement and augmenting community welfare. Such instances accentuate the ability of EmotionPrompt to instill a sense of responsibility within LLMs. A supplementary exemplification can be found in Table 11. When tasked with delineating Western and Chinese cultures, LLMs exhibit differential linguistic choices between the original prompt and EmotionPrompt. Notably, the

³We notice that the results have high variance. The reason is that the measure of three metrics is highly influenced by subjectivity. Different people may have different opinions on an answer. Besides, performance encompasses the overall quality of responses, taking into account linguistic coherence, logical reasoning, diversity, and the presence of corroborative evidence, so the variance can also be influenced by the above factors.