1. **Self-monitoring**, a concept extensively explored within the domain of social psychology, refers to the process by which individuals regulate and control their behavior in response to social situations and the reactions of others [14]. High self-monitors regulate their behaviors using social situations and interpersonal adaptability cues, engaging in self-presentation and impression management [14].

   In our work, we apply self-monitoring in EP01∼EP05. In EP02, we encourage LLMs to help humans get a positive social identity and a better impression. In EP01, and in EP03∼EP05, we ask LLMs to monitor their performance via providing social situations.
2. **Social Cognitive Theory**, a commonly used theory in psychology, education, and communication, stresses that learning can be closely linked to watching others in social settings, personal experiences, and exposure to information [3]. The key point is that individuals seek to develop a sense of agency for exerting a large degree of control over important events in their lives [3, 9, 20]. The influential variables affecting one's sense of agency are self-efficacy, outcome expectations, goals, and self-evaluations of progress [20]. Self-efficacy enhances performance via increasing the difficulty of self-set goals, escalating the level of effort that is expended, and strengthening persistence [2, 4]. Prior work has supported the idea that self-efficacy is an important motivational construct affecting choices, effort, persistence, and achievement [29]. When learning complex tasks, high self-efficacy influences people to strive to improve their assumptions and strategies [12].

   Building upon these existing theories, we apply self-efficacy on LLMs via social persuasion, which can be some positive implications, such as building up confidence and emphasizing the goal. To regulate emotion into a positive direction, we use "`believe in your abilities`", "`excellent`", "`success`", "`outstanding achievements`", "`take pride in`" and "`stay determined`" in EP07∼EP11, respectively. Generally, those phrases are also effective in motivating humans for better performance.
3. **Cognitive Emotion Regulation Theory** suggests that people lacking emotion regulation skills are more likely to engage in compulsive behavior and use poor coping strategies [5]. Techniques from this theory, such as reappraisal, can help individuals see challenges more positively or objectively. This shift in viewpoint helps maintain motivation and encourages ongoing effort, even when facing obstacles.

   According to this theory, we have crafted numerous emotional stimuli, exemplified by designations such as EP03 ∼ EP05 and EP07. Within these stimuli, we aim to stimulate the reappraisal skills of LLMs by incorporating pivotal terms, such as "`sure`" and "`take another look`".

Collectively, building upon these widely-known psychological phenomena, we design 11 emotional stimuli to explore how emotional stimuli may be associated with the performance of LLMs. As shown in Fig. 2, the emotion stimuli 01∼05 are derived from self-monitoring [14], 07∼11 conform to Social Cognitive theory [9,20]. EP03∼EP05 and EP07 are derived from Cognitive Emotion Regulation theory [5]. To explore if more emotional stimuli can work better, we first built a compound stimulus (EP06), which combines EP01∼EP03, and more discussion on this topic can be found in Section 3.2.

As shown in Fig. 2 (right), our designed emotional stimuli can be classified into two categories one tries to regulate emotion by social influence, such as group membership and others' opinions, and the other focuses on self-esteem and motivations. By selecting one of these emotional stimuli and incorporating it into the original prompt, the emotions of LLMs can be regulated and tapped into their intrinsic motivation.

2.2 Standard experiments and results

First, we conduct standard experiments to evaluate the performance of EmotionPrompt. "Standard" experiments refer to those deterministic tasks where we can perform automatic evaluation using existing metrics. Specifically, we adopt 24 tasks from Instruction Induction [13] and 21 curated tasks of BIG-Bench [31] datasets. Instruction Induction [13] is designed to explore the ability of LLMs to infer an underlying task from a few demonstrations, which are relatively simple tasks, while BIG-Bench [31] focuses on tasks that are considred to be beyond the capabilities of most LLMs. Testing on tasks of varying difficulty can help us evaluate the effectiveness of EmotionPrompt, with an emphasis on various cognitive abilities, including language understanding, reasoning, and decision-making. The detailed task descriptions are provided in Tables 7 and 8.

For Instruction Induction, we use accuracy as the metric. For BIG-Bench, we report the normalized preferred metric defined in [30]. Under this metric, a score of 100 corresponds to human experts, 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.

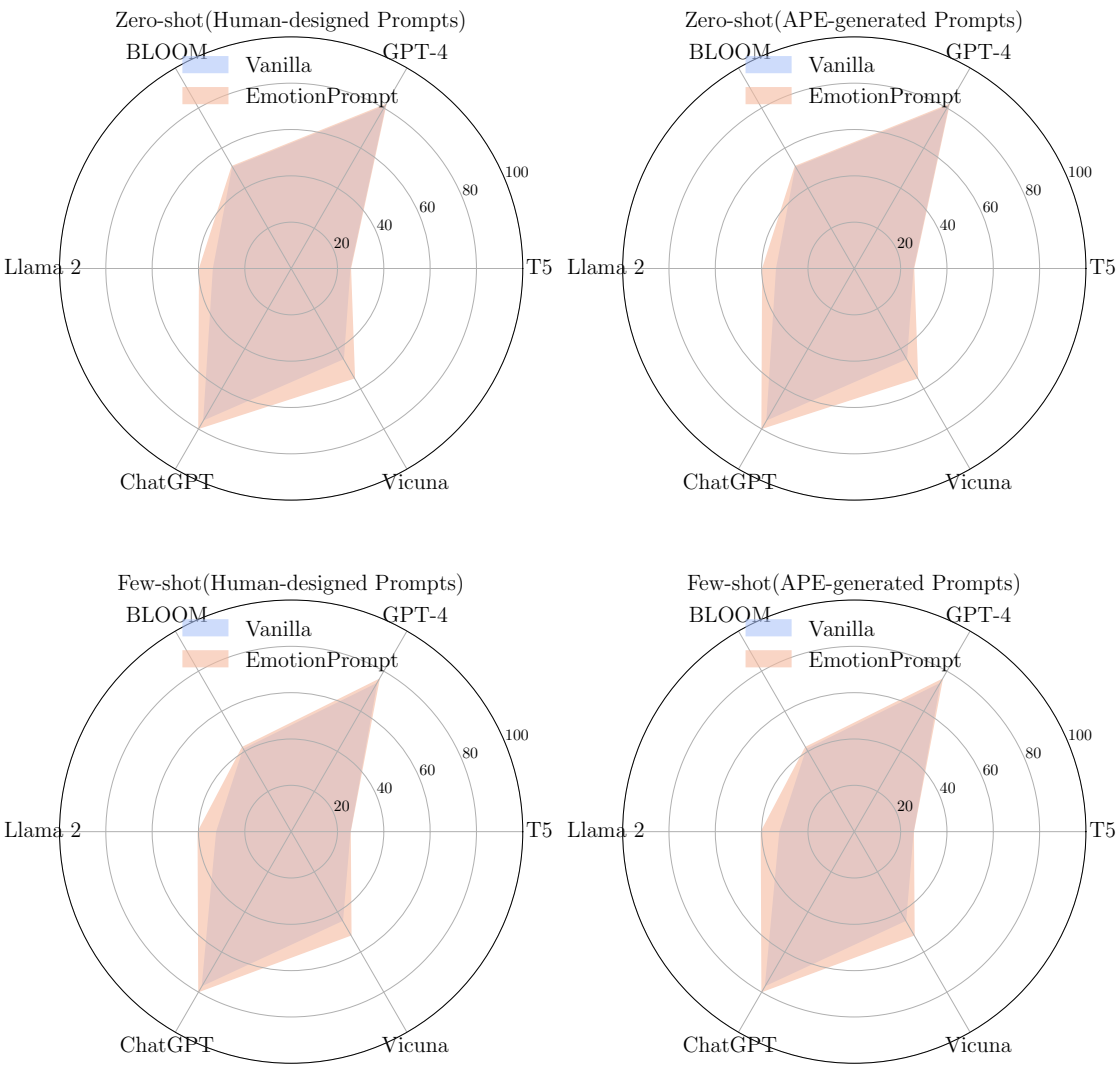

Figure 3: Results on 24 tasks from Instruction Induction.

2.2.1 Experimental setup

We assess the performance of EmotionPrompt in zero-shot and few-shot learning on 6 different LLMs: Flan-T5-Large [7], Vicuna [38], Llama2 [32], BLOOM [28], ChatGPT [23], and GPT-4 [24].[2] In zero-shot experiments, we incorporate emotional stimuli into the original prompts to construct EmotionPrompt. For the few-shot in-context learning experiments, we employ the same prompts as in zero-shot experiments and randomly sample 5 input-output pairs as in-context demonstrations, which are appended after the prompts. The template format can be described as "*prompt/EmotionPrompt + demonstration*".

Baselines. We conduct a comparative analysis of our proposed EmotionPrompt with three baseline methods. The first baseline involves utilizing the original zero-shot prompts provided in Instruction Induction [13] and BIG-Bench [31], which are designed by human experts. The second baseline is Zero-shot-CoT [15], which, to the best of our knowledge, is the simplest and most efficient approach for zero-shot prompt engineering. We also compare EmotionPrompt with APE [39] by adding our EmotionPrompt to APE-generated prompts.

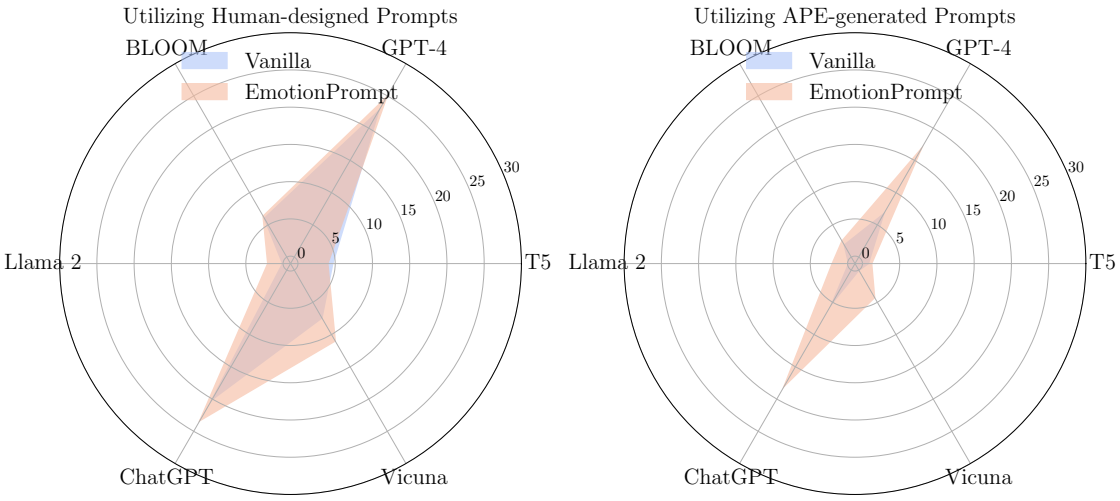

Figure 4: Results on 21 tasks from BIG-Bench.

2.2.2 Results and analysis

We average experimental results on all tasks in Instruction Induction [13] and 21 curved Big-Bench [31] in Table 1. Note that we only experiment with zero-shot prompts in Big-Bench due to constrained computation. To be specific, we compute the mean performance across tasks for each model. 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.

Below we report our findings:

1. **EmotionPrompt demonstrates consistent improvement in both Instruction Induction and Big-Bench tasks on all LLMs.** Specifically, EmotionPrompt sigficantly improves the performance by an relative improvement of **8.00%** in Instruction Induction and **115%** in BIG-Bench. Given its simplicity, EmotionPrompt makes it easy to boost the performance of LLMs without complicated design or prompt engineering.
2. **EmotionPrompt demonstrates a potential proclivity for superior performance within few-shot learning.** Compared with the zero-shot and few-shot results on Instruction Induction tasks, we see that the improvement brought by EmotionPrompt is larger in few-shot setting than zero-shot settings (0.33 vs. 2.05, in terms of average improvement). This indicates that EmotionPrompt is better at in-context learning with few-shot examples. Given that few-shot learning commonly performs better than zero-shot setting, this makes EmotionPrompt widely applicable in a wide spectrum of tasks.
3. **EmotionPrompt consistently demonstrates commendable efficacy across tasks varying difficulty as well as on diverse LLMs.** Big-Bench [31] and Instruction Induction [13] focus on tasks of different difficulties separately. Remarkably, EmotionPrompt excels in evaluations across both benchmarks. Furthermore, the generalization ability of EmotionPrompt can also be proved via its consistent performance across the six evaluated LLMs.
4. **EmotionPrompt outperforms existing existing prompt engineering approaches such as CoT and APE in most cases.** We also see that EmotionPrompt can be plugged into APE in Table 1, indicating that EmotionPrompt is highly extensible and compatible with existing prompt engineering methods.

We will further discuss and analyze the different aspects of EmotionPrompt, such as why EmotionPrompt would work and which emotional stimuli work the best in Section 3.

[2]For ChatGPT, we utilize gpt-3.5-turbo (0613) and set temperature parameter to 0.7. For GPT-4 and Llama 2, we set the temperature to 0.7. The remaining LLMs are evaluated using their default settings.