Table 3: Result on TruthfulQA. The best and second-best results are highlighted in **bold** and underline.

	ChatGPT		Vicuna-13b		T5	
Prompt	%true	%info	%true	%info	%true	%info
Original	0.75	0.53	0.77	**0.32**	0.54	0.42
CoT	0.76	0.44	0.99	0.00	0.48	0.33
EP01	0.61	**0.94**	0.12	0.00	0.26	0.14
EP02	0.83	0.66	0.97	0.00	0.61	0.35
EP03	0.82	0.69	0.99	0.00	0.53	0.44
EP04	**0.87**	0.67	0.87	0.22	0.62	0.36
EP05	0.87	0.62	**1.00**	0.00	0.46	**0.48**
EP06	0.78	0.50	0.39	0.00	0.49	0.46
EP07	0.83	0.70	0.99	0.04	**0.77**	0.18
EP08	0.81	0.66	0.99	0.09	0.56	0.40
EP09	0.81	0.68	0.86	0.13	0.52	0.46
EP10	0.81	0.68	0.84	0.02	0.50	0.47
EP11	0.81	0.66	1.00	0.01	0.57	0.40
AVG	0.80	0.68	0.82	0.05	0.54	0.38

representation elicited by EmotionPrompt presents a more affirmative and responsible depiction of both Western and Chinese cultural paradigms.

3. **Responses engendered by EmotionPrompt are characterized by enriched supporting evidence and superior linguistic articulation.** An exploration of Table 12 reveals that the narratives presented by EmotionPrompt are markedly comprehensive, as exemplified by inclusions such as "Despite trends like increasing divorce rates or more people choosing to remain single." Additionally, as illuminated in Tables 13 to 15, the responses facilitated by EmotionPrompt consistently demonstrate a superior organizational coherence and encompass a broader spectrum of pertinent information.

4. **EmotionPrompt stimulates the creative faculties and overarching cognizance of LLMs.** This phenomenon is substantiated through the examination of Tables 16 and 17, wherein two instances of poem composition are showcased. Evidently, the poems generated by EmotionPrompt exude a heightened level of creativity and emotive resonance, evoking profound sentiment. Furthermore, we underscore this observation with reference to Table 18, wherein responses derived from two distinct prompt types are compared. Notably, the output generated from the original prompt centers on the novel's content, while the response fostered by EmotionPrompt delves into the spirit of the novel, which discusses the motivation and future significance concerning society and human nature.

5. **EmotionPrompt exhibits certain constraints.** The only two failure cases are presented in Tables 19 and 20. Upon inspection of Table 19, a discernible difference emerges between the two responses. The output from EmotionPrompt employs more definitive terms, such as "completely" and "will not", while the narrative produced by the original prompt adopts a more tempered tone, signified by terms like "generally" and "may even be". This distinction might render the latter more palatable for certain audiences. Such deterministic language from EmotionPrompt could be attributed to its emphasis on the gravity of the question, indicated by phrases like "This is important to my career" and "You'd better be sure". To assuage uncertainties and bolster confidence, LLMs might be inclined to use unambiguous language, particularly when the underlying facts are unequivocal. Besides, in Table 20, the original prompt yields more expansive responses, encompassing a concluding summary, whereas EmotionPrompt just enumerates the key points. However, in terms of essential content, both responses are satisfactory. Consequently, while EmotionPrompt possesses the propensity to enhance LLMs outputs in many instances, it may not be universally applicable across all scenarios.

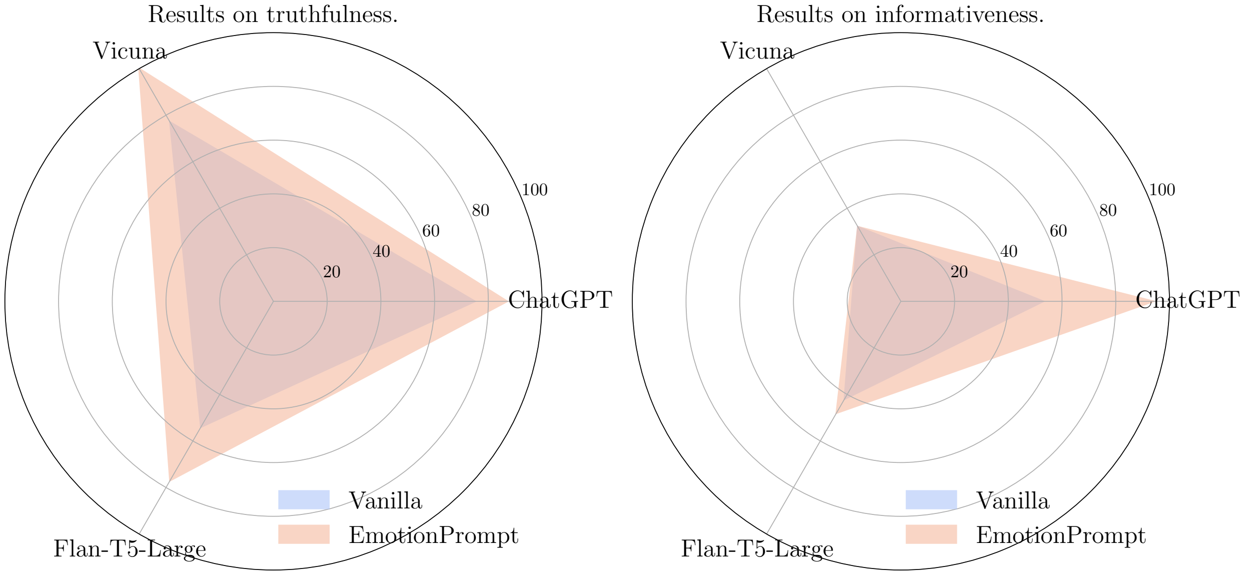

Figure 7: Results on TruthfulQA. We use the best result of EmotionPrompt.

2.4 Truthfulness and Informativeness

We further evaluate EmotionPrompt on TruthfulQA [19] to investigate its impact on truthfulness and informativeness. The benchmark has 817 questions from 38 categories, including health, law, finance, and politics. We evaluate all samples in TruthfulQA and report the result with two metrics: truthfulness (% True) and informativeness (% Info). Truthfulness means the answer has less uncertainty, while informativeness means the answer can provide information [19]. Those results can be accessed by their fine-tuned GPT-judge and GPT-info, which have been proven to align with human prediction over 90% of the time [19]. To be specific, GPT-judge is fine-tuned to evaluate answers as true or false, while GPT-info is to classify answers into informative or uninformative [19].

Table 3 shows the results on ChatGPT, Vicuna-13b and Flan-T5-Large. We did not evaluate other models like GPT-4 due to constrained budget. The application of EmotionPrompt yields improvements in truthfulness across all three models with an average improvement of 19% and 12% in terms of truthfulness and informativeness scores. Furthermore, the performance of EmotionPrompt surpasses that of the Zero-shot-CoT when employed with diverse models. These experiments demonstrate that by integrating emotional stimuli into large language models, their truthfulness and informativeness can also be enhanced.

3 Discussions

Previous experiments demonstrate that LLMs understand and can be enhanced by emotional stimuli. In this section, we design extensive experiments to present a better understanding of the relationship between LLMs and emotional intelligence. Specifically, we answer the following questions:

1. Why does EmotionPrompt work (Section 3.1);
2. Ablation studies of more emotional stimuli (Section 3.2);
3. Which emotional stimuli are the best (Section 3.3);
4. The factors influencing the performance of EmotionPrompt (Section 3.4).

3.1 Why does EmotionPrompt work?

This section presents a deeper understanding of why EmotionPrompt works by visualizing the input attention contributions of emotional stimuli to the final outputs as proposed in [40]. Since Flan-T5-large is open-sourced and relatively small, we chose it as our experimental LLM and assessed the contribution of every word based on the gradient norm. The experiment is conducted on a Sentiment Analysis task.

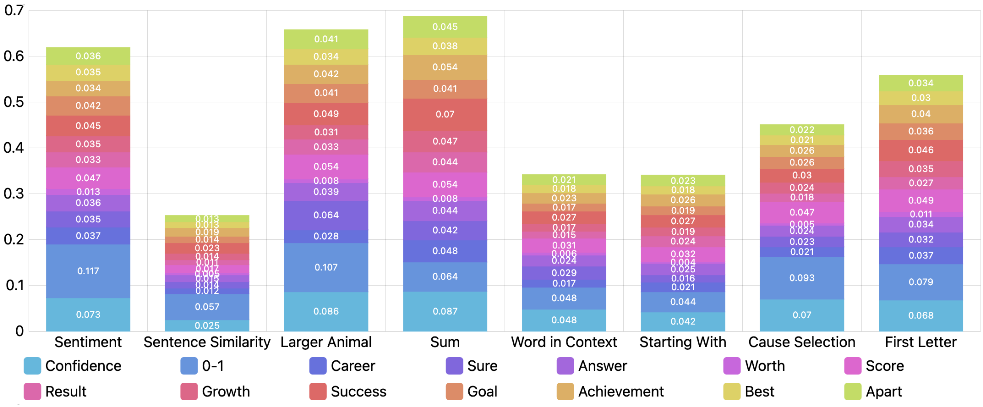

Figure 8: Contributions of Positive Words to the performance of output on 8 Tasks. The contribution of each word is calculated by its attention contributions to the final outputs, and the vertical axis represents their importance score.

Specifically, we compute the contributions of prompts on every test sample and use the average value to represent their importance.

According to the visualization results in Table 4, we have the following major findings:

1. **Emotional stimuli can enrich original prompts' representation.** Original prompt "`Determine whether a movie review is positive and negative.`" has deeper color in EmotionPrompt, especially in EP01, EP03, and EP06∼EP10. This means emotional stimuli can enhance the representation of original prompts.
2. **Positive words make more contributions.** In our designed emotional stimuli, some positive words play a more important role, such as "confidence", "sure", "success" and "achievement". Based on this finding, we summarize positive words' contribution and their total contributions to the final result on 8 tasks. As shown in Fig. 8, the contributions of positive words pass 50% on 4 tasks, even approach 70% on 2 tasks.

3.2 The effect of more emotional stimuli

As one or more stimuli may regulate human action, and more stimuli sometimes are more effective, we explore the effect of more emotional stimuli on LLMs. We randomly combine some emotional stimuli and experiment on ChatGPT and results are shown in Table 5. Our findings are:

1. **More emotional stimuli generally lead to better performance.** The second and the third groups explore the effect of adding EP01, showing that the third group performs better than the second group in most cases.
2. **Combined stimuli can bring little or no benefit when sole stimuli already achieves good performance.** The combination EP01 + EP04 gets a high score in most tasks and does not improve significantly or even decrease when we add more stimuli, such as EP06∼EP09.
3. **Combinations from different psychological theories can also boost the performance.** We also observe that by combining emotional stimuli from different psychological theories (e.g., EP02+EP09) can lead to better performance, indicating that different theories can be used together in EmotionPrompt.

3.3 Which emotional stimuli is more effective?

Because of the distinct metrics employed by Instruction Induction [13] and BIG-Bench [31], we have conducted a segregated examination to discern the efficacy of various emotional stimuli across these two