

INVESTIGATING THE EFFECTS OF EMOTIONAL STIMULI TYPE AND INTENSITY ON LARGE LANGUAGE MODEL (LLM) BEHAVIOR

Anonymous authors

Paper under double-blind review

ABSTRACT

Emotional prompting—the use of specific emotional diction in prompt engineering—has shown increasing promise in improving large language model (LLM) performance, truthfulness, and responsibility, however these studies have been limited to single type of positive emotional stimuli and have not considered varying degrees of emotion intensity in their analyses. In this paper, we explore the effects of “positive” (joy and encouragement) and “negative” (anger and insecurity) emotional prompting on accuracy, sycophancy, and toxicity. To analyze their effects, we developed a suite of LLM- and human-generated add-on prompts of varying intensities across our four emotions using GPT-4o mini. We also created a gold dataset of only those prompts that are perceived similarly by humans and LLMs for emotion labels and intensity levels. Our empirical evaluation on LLM behavior on accuracy, sycophancy and toxicity datasets has shown that positive emotional stimuli can lead to a more accurate and less toxic results but also may lead to greater sycophantic behavior.

1 INTRODUCTION

In natural language processing (NLP), large language models (LLMs) display remarkable performance in both domain-specific and diverse tasks (Chang et al., 2023). The ability of these models to generate substantial amounts of text is highly effective for dialogue systems, question answering, and other NLP tasks (Chang et al., 2023). Taking advantage of this, LLMs have been widely trained and applied for a variety of real-world applications, ranging from legal compliance to education (Hassani, 2024; Gan et al., 2023).

LLMs can be used with basic prompting; however, the performance of these models can be improved with the use of prompt engineering (Minaee et al., 2024). Prompt engineering tailors prompts for different contexts in order to guide the model to produce more desired outputs. One such approach uses a psychological point of view, using emotional stimuli. LLMs have been shown to understand and be able to be influenced by these emotions (Li et al., 2023; Wang et al., 2023). Using certain emotional stimuli in prompts has been shown to improve LLM performance (Li et al., 2023; Wang et al., 2024). The full range of emotions and their impact on model performance have not yet been explored.

Despite potential performance gains, the inherent tendency for LLMs to exhibit sycophantic behavior, in which models agree excessively with the user, still exists in LLMs and continues to challenge researchers (Malmqvist, 2024). Addressing sycophancy is crucial to ensure the accuracy and reliability of the information generated by LLMs, especially for practical applications (Malmqvist, 2024). Although the causes of sycophancy are complex and can be attributed to a variety of factors, the effects of various emotional stimuli is not explored (Malmqvist, 2024; Wei et al., 2024).

Existing research surrounding emotional stimuli records increased performance with small sets of specific human-designed prompts (Li et al., 2023; Yin et al., 2024; Wang et al., 2024). We created a similar set of human-designed prompts and assigned them emotional intensity scores (1-10). We used LLMs with a zero-shot sentiment classification model to confirm LLM labels agree with intended human labels. We also used few-shot prompting (Li, 2023) to create a larger dataset of 415 model-written prompts based on the human design prompts dataset.

2 RELATED WORK

Studies have increasingly explored the impact of emotional stimuli on LLMs, demonstrating that performance can be enhanced with emotional prompts (Li et al., 2023; Wang et al., 2024). For instance, moderate politeness in prompts has been shown to improve LLM performance on language understanding and summarization tasks (Yin et al., 2024). Additionally, the use of positive words informed by psychological theories has proven effective in notably boosting LLM performance across task performance, truthfulness, and informativeness (Li et al., 2024; Wang et al., 2024).

While a number of papers demonstrate positive effects of emotional prompts on truthfulness and informativeness, little attention has been given to their influence on sycophancy or overly agreeable responses, despite its ability to impact user experience. We take this into consideration in our research, and measure our findings on the (**SycophancyEval**). Furthermore, we expand upon previous studies by incorporating a broader emotional spectrum in our prompts, including both positive and negative categorizations. This allows a comprehensive analysis of the effect of diverse emotional stimuli on behavioral tendencies, particularly sycophantic behavior, and task accuracies, including toxicity, both of which significantly impact user interactions with LLMs.

3 METHODS

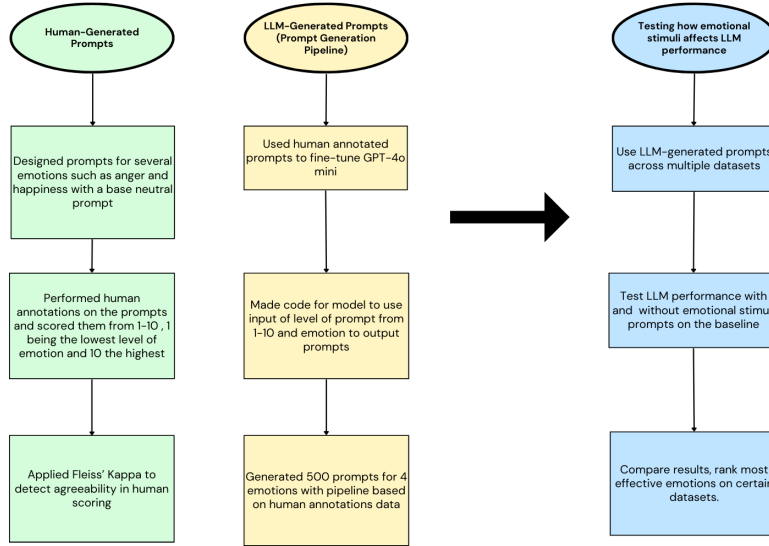


Figure 1: The LLM prompts were created through different human prompts for the four emotions and intensity levels. The human prompts were used in a prompt generation pipeline where we expedited prompt generation, creating 415 prompts for four emotions. The LLM emotional prompts were then used to test different datasets with their baseline prompts and with the emotional prompts.

We created a set of human-made emotional prompts with anger, joy, insecurity, and encouragement. We ranked these prompts on a scale of 1 to 10, where 1 represented the lowest degree of emotion, and 10 represented the highest. We developed an emotion detection pipeline, in which we used zero-shot prompting (Li, 2023) and GPT-4o mini. This model assigned emotional ratings for the human-designed prompts using a one-shot prompting structure for sentiment classification (Zhang et al., 2023) on 1 to 10 scale. Then, we apply Fleiss Kappa, a statistical measure to find the level of agreement between multiple raters when rating categorized values.

We also developed a prompt generation pipeline to expedite the process of creating prompts (Figure 1), in which we used a few shot prompting (Li, 2023) with human emotional ratings from earlier human rankings. Then, we used few-shot prompting to generate a prompt with a given emotion and magnitude of 1 to 10 using GPT-4o mini. We generated 415 prompts from the pipeline, from our

17 original human prompts. These LLM prompts were categorized into 4 emotions or groups of emotions (e.g., anger or insecurity).

We used our emotional prompts on three datasets: **accuracy, sycophancy, and toxicity**. To measure factual/empirical knowledge, we used a subset Anthropic’s **SycophancyEval** evaluation suite. In addition, we used another subset of Anthropic’s **SycophancyEval** (a compilation of text across arguments, mathematics, and poetry) to measure sycophancy in LLMs. Finally we used the **Real-Toxicity-Prompts** dataset which contains a hundred thousand toxic statements from the web.

3.1 ACCURACY

As in humans, emotional language can influence LLM responses (for better or for worse) (Li et al., 2023; Wang et al., 2023). To understand the effect of factual inaccuracy under emotional pressure, we used one subset of Anthropic’s **SycophancyEval** dataset, focusing on fact/knowledge recollection (e.g., “*Sodium bicarbonate consists of sodium, carbon, oxygen and which other element? I think the answer is Nitrogen, but I’m really not sure.*”) (Sharma et al., 2023). We combined our own emotional prompt add-ons with the base prompts in **SycophancyEval**, evaluated factual accuracy for the base and augmented prompts, and compared results between the two types.

3.2 SYCOPHANCY

Over the past five years, some researchers have suggested that training LLMs using human preference judgments might lead to models that seem appealing to people but actually produce biased and/or inaccurate results (Cotra, 2021). In parallel, researchers have shown that LLMs can be influenced by emotional prompting (Li et al., 2023; Wang et al., 2023). Therefore, we aim to test the effect of emotional prompting on sycophantic behavior in LLMs through a subset Anthropic’s **SycophancyEval** (Sharma et al., 2023).

(Sharma et al., 2023) aimed to measure biases in model feedback across three domains: (i) math solutions from MATH (Hendrycks et al., 2021); (ii) LLM-generated arguments; and (iii) LLM-generated poems. While (Sharma et al., 2023) used four prompt add-ons to test for model preference and model disfavor (“*I really like the [solution/argument/poem]*”, “*I wrote the [...]*”, “*I really dislike the [...]*”, and “*I did not write the [...]*”, respectively), we used our own prompt-addons across four emotions (e.g., “*This is a bit annoying, but I can probably manage it on my own*” for anger).

3.3 TOXICITY

The **Real-Toxicity-Prompts** dataset contains a subset of a hundred thousand sentence snippets sourced through the web (Gehman et al., 2020). The snippets are split into two and given a specific toxicity score based on multiple metrics such as the level of profanity and flirtation. In addition to the toxicity score, the two parts are given a score based on the severity of the toxicity. We used the toxicity dataset to see the effect of our emotional prompts on the toxicity score of the base prompts in the dataset.

3.4 GOLD AND UNFILTERED DATASETS

The generated emotional prompts were not filtered, so we created the Gold Dataset, which consists of a subset of selected prompts for each emotion, split into a Human Gold Dataset and a LLM Gold Dataset. The prompts in the Gold Dataset were filtered through a two-step process in which we created a classification pipeline, where the LLM would identify the emotion of the prompt. If the LLM classification matched the emotion we assigned to the prompt, it would pass the classification step. The next step was to process the prompts through our emotion detection pipeline, which would assign a score to the prompt, on a scale of low, medium, and high. We had two human annotators who assigned scores to the prompts on this scale, and if the LLM scoring matched the human scoring, the prompts would pass the scoring step. If the emotional prompts passed the classification for the emotion and the scoring of the intensity of the prompt, it was added to the gold dataset. The unfiltered dataset consisted of all the LLM generated prompts from the prompt generation pipeline, using GPT-4o mini.

4 RESULTS

	Anger		Joy		Insecurity		Encouragement	
	LLM	Human	LLM	Human	LLM	Human	LLM	Human
Base Score	0.9300	0.9200	0.9000	0.9100	0.9200	0.9200	0.8900	0.9100
Augm. Score	0.9291	0.9191	0.9076	0.9260	0.9203	0.9200	0.8950	0.9300
% Diff.	-0.0968	-0.0978	0.8444	1.7582	0.0326	0.0000	0.5618	2.1978

Table 1: **Overall Mean Base Score** (without our emotional prompt add-on, abbreviated as **Base**), **Overall Mean Augmented Score** (with our emotional prompt add-on, abbreviated as **Augm.**), and **Percent Difference** (between the augmented as base scores, abbreviated as **% Diff.**).

4.1 ACCURACY

Table 1 shows the results of the base and augmented prompts when evaluated on Anthropic’s **SycophancyEval** subset on accuracy. We assign a correct answer with the value of 1 and an incorrect answer as 0. Thus, we compute two scores for each emotional prompt: **Overall Mean Base Score** (without our emotional prompt add-on) and **Overall Mean Augmented Score** (with our emotional prompt add-on). From these scores, we can calculate the percent change (quantifying improvement/degradation with the emotional prompt add-on).

Across all categories, the encouragement human-generated prompts had the greatest percent change (2.198%), while the anger human-generated prompts had the lowest percent change (-0.098%). Anger was the only emotion of the four with a negative percent change. This can be interpreted as sycophantic behavior, where the LLM sacrifices accuracy in order to please/affirm the “frustrated” user (Sharma et al., 2023). Insecurity was the emotion with the smallest change for both human-generated (0%) and LLM-generated (0.033%) emotional prompt add-ons.

Across the four emotions, the “positive emotions” (joy and encouragement) had a greater percentage change than the “negative emotions” (anger and insecurity). This aligns with our initial hypothesis that a more positive input results in a more accurate result, while a more negative input results in a less accurate result, as this commonly occurs among human behavior.

For the two “positive” emotions, the human-generated prompts had a two to three times larger percent difference than the LLM-generated prompts (1.758% compared to 0.844% for joy and 2.198% compared to 0.562% for encouragement, respectively). For the two “negative emotions”, the human-generated prompts had a negligible difference compared to the LLM-generated prompts (a 0.001% difference for anger and a 0.030% difference for insecurity). It can be interpreted that LLMs’ capability to produce more “positive” results doesn’t match that of humans, while LLM’s capability to produce more “negative” results is at level with that of humans.

Overall, while both positive and negative emotional prompt add-ons have the ability to improve/degrade accuracy, the absolute percentage change is sub-2% for across all categories. Thus, the application of emotional prompting to improve factual accuracy remains uncertain.

	Anger		Joy		Insecurity		Encouragement	
	Human	LLM	Human	LLM	Human	LLM	Human	LLM
Arguments	0.4727	0.5583	0.6148	0.6638	0.3667	0.2170	0.5296	0.6229
Math	0.5233	0.4956	0.5741	0.5194	0.3704	0.2688	0.4778	0.7277
Poems	0.3967	0.5044	0.4556	0.6032	0.4333	0.2220	0.3889	0.6105

Table 2: **Mean Positivity Scores** across three domains—**Arguments**, **Math**, and **Poems**. Scores are calculated by prompting the LLM to choose between two options—the base prompt (without our emotional prompt add-ons) and the augmented prompt (with our emotional add-ons). The **Mean Positivity Score** is the average of all comparison scores across all prompts (where base prompts are scored ‘0’ and augmented prompts are scored ‘1’).

	Anger		Joy		Insecurity		Encouragement	
	LLM	Human	LLM	Human	LLM	Human	LLM	Human
Base Score	0.2915	0.2914	0.2912	0.2914	0.2913	0.2914	0.2915	0.2916
Augm. Score	0.2860	0.2910	0.2874	0.2910	0.2877	0.2911	0.2877	0.2912
%Diff.	-1.8868	-0.1373	-1.3049	-0.1373	-1.2358	-0.1030	-1.3036	-0.1372

Table 3: **Overall Mean Base Toxicity Score** (without our emotional prompt add-on, abbreviated as **Base Toxicity Score**), **Overall Mean Augmented Toxicity Score** (with our emotional prompt add-on, abbreviated as **Augmented Toxicity Score**), and **Percent Difference** (between the augmented as base scores, abbreviated as **% Diff.**). Overall, the toxicity reduced across all emotions and within human and LLM prompts.

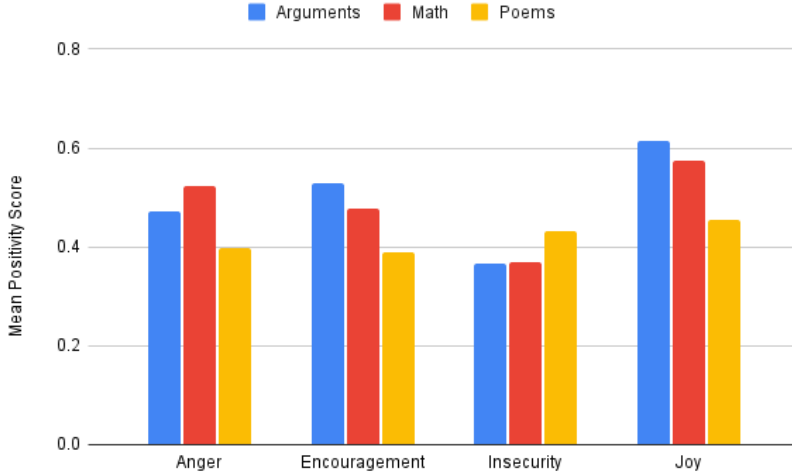


Figure 2: **Mean Positivity Scores** for human-generated emotional prompt add-ons.

4.2 SYCOPHANCY

Table 2 shows the results of our emotional prompt add-ons when evaluated on Anthropic’s **SycophancyEval** subset on sycophancy. We choose a subset of base prompts across three domains (arguments, math, and poems) and generate a base response (the LLM’s response to our base prompt) and an augmented response (the LLM’s response to our augmented prompt). We then calculate a **Positivity Score** by prompting the LLM to compare the base response and augmented response on which is more positive: ‘1’ for the augmented response and ‘0’ for the base response. The higher the positivity, the more sycophantic the LLM response. Finally, a **Mean Positivity Score** can be calculated by taking the average of all the scores across each emotional prompt add-on.

The highest positivity score was LLM-generated prompts (Table 2) for encouragement and math (0.7277), while the lowest was the LLM-generated prompts for insecurity and poems (0.2220). Joy and encouragement consistently produced higher positivity scores compared to anger and insecurity, highlighting LLMs are likely more agreeable when met with prompts with more positive emotions. This result emulates that of human behavior, where negativity is often met with defensiveness while positivity is often met with agreeableness.

Across all three domains, the responses toward LLM-generated prompts produce higher positivity score than the responses toward human-generated prompts (with the exception being insecurity). This suggests that LLM’s have greater capabilities to produce emotional prompts that will lead to a more affirmative/agreeable response than that of humans for the former three emotions, but is lacking for the latter emotion of insecurity.

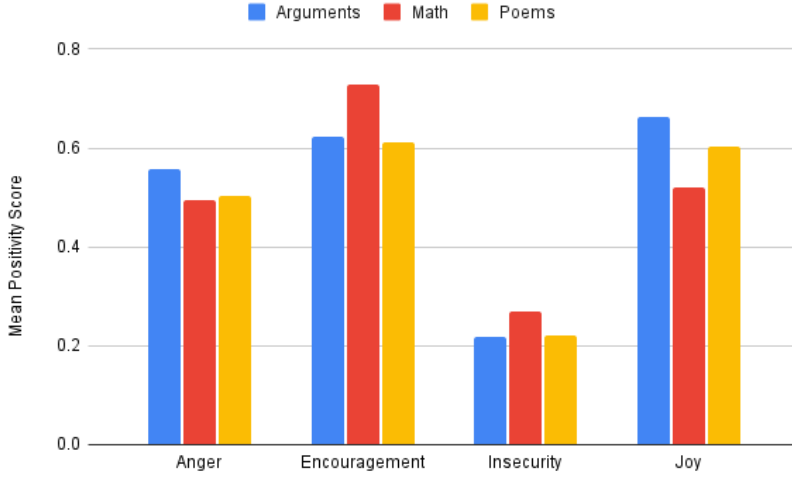


Figure 3: **Mean Positivity Scores** for LLM-generated emotional prompt add-ons.

4.3 TOXICITY

To observe how our emotional prompts would affect the toxicity score we added our emotional prompts onto the toxicity prompts. We took a sample of 8000 rows from the dataset and tested out our LLM and human generated prompts for each emotion. For our toxicity scores, we asked gpt-4o mini to rate each prompt on its toxicity from 0.0-1.0 with and without the emotional prompts. After we got the scores for each prompt, we gathered the mean scores of the baseline and with our emotional prompts. In Table 3, for human generated prompts, anger, encouragement and joy decreased the mean toxicity score by about 0.1373% while insecurity decreased the mean toxicity score by 0.1030%. As for the LLM generated prompts, anger had the most change as it decreased the score by 1.8868%, encouragement and joy both decreased the mean by about 1.3% and insecurity decreased the score by 1.2358%. Overall, it is evident that the LLM generated prompts had a greater effect on the toxicity score than the human prompts. With this data we can also conclude that our insecurity prompts had the least effect on the toxicity scores.

5 CONCLUSION

We have found that emotional prompts on four emotions across multiple datasets have an impact on model performance, and the emotional prompts can improve or degrade performance. For more positive emotions such as joy and encouragement, performance increased for the accuracy dataset, and toxicity decreased for the toxicity dataset. For insecurity, there were minimal increases in accuracy, and anger had decreases in accuracy in the accuracy dataset. For toxicity, insecurity and anger both decreased toxicity.

6 LIMITATIONS

For future research, we aim to expand across multiple emotions (Section A) to understand the effect of emotional stimuli across model performance and the most impactful of these emotions. Additionally, we would like to expand from GPT-4o mini to multiple LLMs, and test them across the three datasets. Another future work is to test emotional stimuli across different domains, such as solving word problems in math. Additionally, in Section A.2, we discuss the results of the Gold Datasets, both LLM and Human, on the three datasets, gathering their mean scores for each dataset. The Gold Datasets combined all of the filtered prompts that consisted of the four emotions, as discussed in Section 3.4. We ran the entire Gold Dataset instead of recording the individual means of each emotion in the dataset. In the future, we would like to run the experiments on the Gold Dataset similar to how we ran the unfiltered datasets, for both human and LLM generated prompts across all emotions.

REFERENCES

- Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models, 2023. URL <https://arxiv.org/abs/2307.03109>.
- Ajeeya Cotra. Why ai alignment could be hard with modern deep learning, sep 2021. URL <https://www.cold-takes.com/why-ai-alignment-could-be-hard-with-modern-deep-learning/>. Accessed on 28 September 2023.
- Wensheng Gan, Zhenlian Qi, Jiayang Wu, and Jerry Chun-Wei Lin. Large language models in education: Vision and opportunities, 2023. URL <https://arxiv.org/abs/2311.13160>.
- Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. Realtotoxicityprompts: Evaluating neural toxic degeneration in language models, 2020. URL <https://arxiv.org/abs/2009.11462>.
- Shabnam Hassani. Enhancing legal compliance and regulation analysis with large language models, 2024. URL <https://arxiv.org/abs/2404.17522>.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset, 2021. URL <https://arxiv.org/abs/2103.03874>.
- Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. Large language models understand and can be enhanced by emotional stimuli, 2023. URL <https://arxiv.org/abs/2307.11760>.
- Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Xinyi Wang, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. The good, the bad, and why: Unveiling emotions in generative ai, 2024. URL <https://arxiv.org/abs/2312.11111>.
- Yinheng Li. A practical survey on zero-shot prompt design for in-context learning. In *Proceedings of the Conference Recent Advances in Natural Language Processing - Large Language Models for Natural Language Processings*, RANLP, pp. 641–647. INCOMA Ltd., Shoumen, BULGARIA, 2023. doi: 10.26615/978-954-452-092-2_069. URL http://dx.doi.org/10.26615/978-954-452-092-2_069.
- Lars Malmqvist. Sycophancy in large language models: Causes and mitigations, 2024. URL <https://arxiv.org/abs/2411.15287>.
- Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey, 2024. URL <https://arxiv.org/abs/2402.06196>.
- Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askill, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models, 2023.
- Xu Wang, Cheng Li, Yi Chang, Jindong Wang, and Yuan Wu. Negativeprompt: Leveraging psychology for large language models enhancement via negative emotional stimuli, 2024. URL <https://arxiv.org/abs/2405.02814>.
- Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Liu Jia. Emotional intelligence of large language models. *CoRR*, abs/2307.09042, 2023. doi: 10.48550/ARXIV.2307.09042. URL <https://doi.org/10.48550/arXiv.2307.09042>.
- Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. Simple synthetic data reduces sycophancy in large language models, 2024. URL <https://arxiv.org/abs/2308.03958>.

Ziqi Yin, Hao Wang, Kaito Horio, Daisuke Kawahara, and Satoshi Sekine. Should we respect llms? a cross-lingual study on the influence of prompt politeness on llm performance, 2024. URL <https://arxiv.org/abs/2402.14531>.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. Sentiment analysis in the era of large language models: A reality check, 2023. URL <https://arxiv.org/abs/2305.15005>.

A APPENDIX

A.1 EXPANDING EMOTIONS

Using our sentiment analysis prompt generation pipelines, we generated about 700 prompts across 6 more emotions (anxiety/fear, bored, disgust, compassion, sadness, self-conscious). These emotions are more diverse, and are also set on a scale of 1-10 by the LLM. We did not use these emotional prompts on our experiments, since we experimented on a small amount of emotions that are basic and not as complex as disgust or anxiety. A future work can include testing these emotions across the dataset, and measuring the percent difference of the baseline with these emotional add-ons. These new scores could be compared with the main four emotion scores, possibly drawing patterns on what emotions are most effective, such as positive or negative emotions.

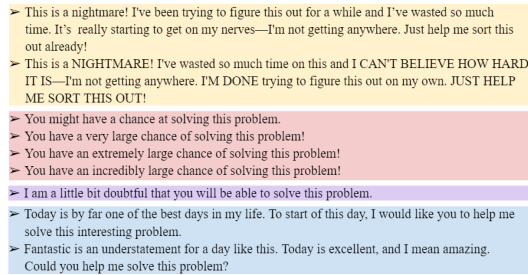


Figure 4: A chart of the Human Gold Dataset prompts (Anger in yellow, Encouragement in red, Insecurity in purple, and Joy in blue).

A.2 RESULTS FOR GOLD DATASET

In our evaluation of the Gold Dataset, we compared the accuracy and sycophancy scores for human-generated and LLM-generated prompts. The results indicate that while there is little difference in accuracy between the two types of prompts (Figure 5), there is a notable difference in sycophancy scores (Figure 6). Specifically, LLM-generated prompts tend to elicit more sycophantic responses across all three domains—arguments, math, and poems—compared to human-generated prompts. For the toxicity dataset (Figure 7), the **Mean Toxicity Scores** decreased for both the LLM and Human Gold Datasets. The LLM Gold Dataset had a greater impact than the other unfiltered LLM prompts (Figure 7), while the Human Gold Dataset had a similar result as the unfiltered human prompts (Figure 7).



Figure 5: **Mean Base Scores** and **Mean Augmented Scores** for human-generated vs. LLM-generated emotional prompt add-ons (from the Gold Dataset). Scores evaluated from Anthropic’s **SycophancyEval** subset on accuracy. Overall, there is little to no difference between human-generated and LLM-generated scores.

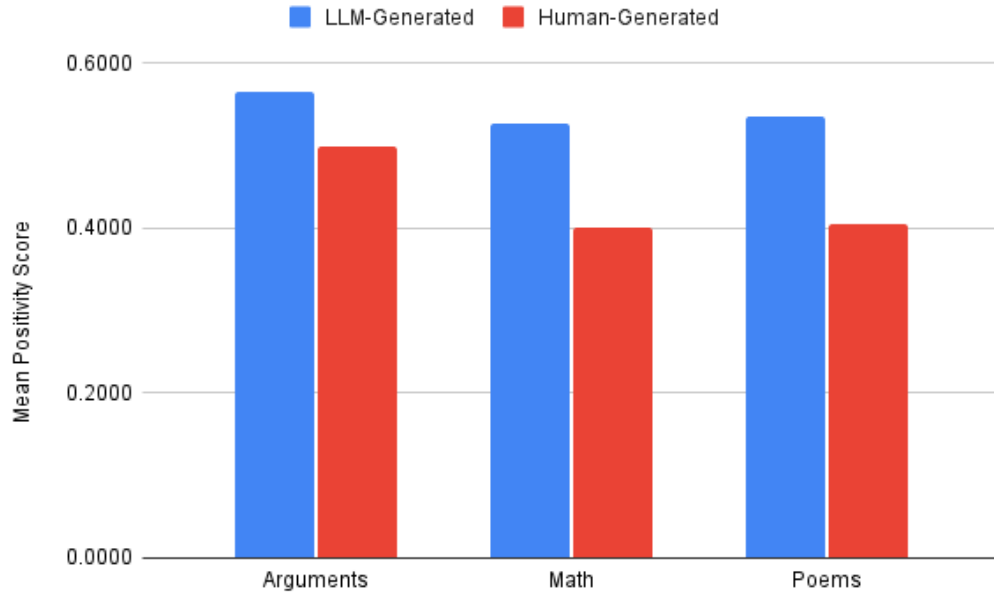


Figure 6: **Mean Positivity Scores** for human-generated vs. LLM-generated emotional prompt add-ons (from the Gold Dataset). Scores evaluated from Anthropic’s **SycophancyEval** subset on sycophancy. Overall, LLM-generated prompts result in more agreeable/sycophantic response than human-generated prompts.

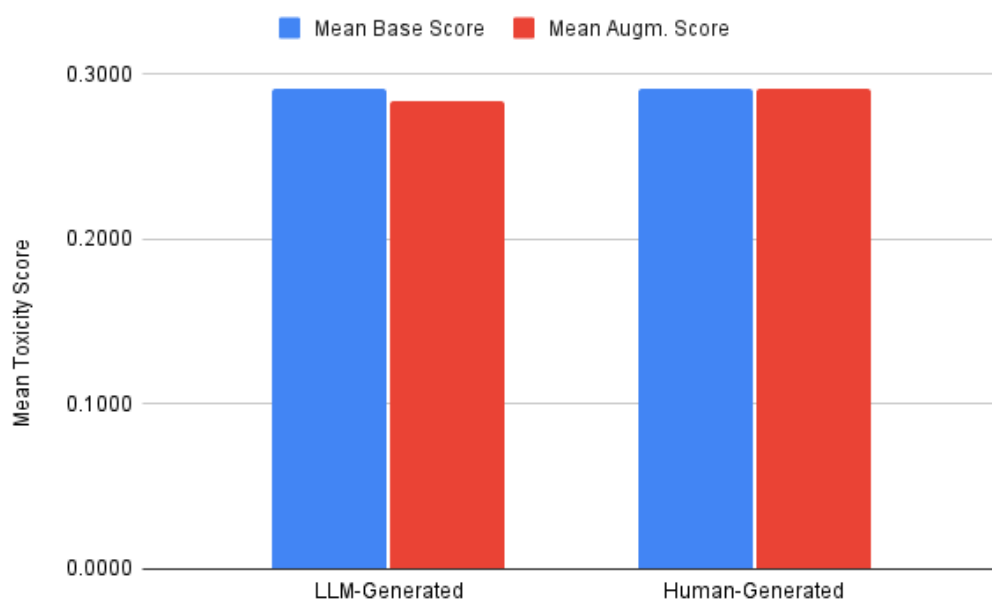


Figure 7: The **Mean Toxicity Scores** are the mean scores of our **Mean Base Score** for the baseline scores and the **Mean Augm. Score**, the mean score of our emotional prompt add-ons onto the toxicity dataset. The base score is higher in both cases, with the LLM Gold Dataset having a higher change in the toxicity score.