LLM-based Affective Text Generation Quality Based on Different Quantization Values

Anonymous ACL submission

Abstract

Large language models in Natural Language Processing (NLP) exhibit a remarkable capacity for tasks such as language generation, translation, and contextual comprehension. In order to achieve these results, the model uses a large number of parameters, requiring a significant number of computational resources for training or utilization. Reducing the precision bits makes the models smaller, resulting in fewer computational resources needed to use the models, at the cost of lowering overall accuracy. This paper addresses the trade-off between different quantization values, GPU RAM utilization, and text quality in affective text generation (e.g., "I really enjoy running in the snow-covered forest"). To evaluate, we use an emotion classifier and ten seed prompts to generate affective text. We test three setups of precision bits (8, 16, and 32) across two open weight language models. Our findings demonstrate that bit reductions leads to memory savings, achieving a reduction of 76 %. However, this optimization comes with a trade-off, leading to a decrease of up to 10 pp in F_1 score for larger models and an increase of 10 pp for smaller models, along with roughly double the inference time. In terms of text quality, larger models at lower quantization levels generally outperform smaller, higher-precision models – while requiring similar memory.

1 Introduction

Natural Language Processing (NLP) has transformed significantly with the emergence of Large Language Models (LLM, Touvron et al., 2023; Radford et al., 2018; Brown et al., 2020; Radford et al., 2019), showcasing unprecedented capabilities in language generation, translation, and contextual comprehension. However, these advancements come with a cost – a significant demand for computational resources. A popular solution is quantization, which involves converting high-precision values to lower-precision ones (e.g., from 16 to 8 bits) and overall model size (Hu et al., 2022; Dettmers et al., 2023), making LLMs consume less GPU memory, require less storage, and be more energy-efficient. However, the primary drawback of quantization is the potential loss of accuracy, with higher bit reductions posing a greater risk. This reduction in precision sets the stage for exploring the trade-off between computational resources, such as GPU RAM, inference time, and text quality. Our research focuses on the impact of quantization on affective text generation. Therefore, we aim to answer: **What is the trade-off between precision in bits on LLMs in affective text generation?**

Post-training quantization is popular due to its simplicity and effectiveness. Jacob et al. (2018) demonstrated that it can reduce full-precision models to 8-bit integers with minimal accuracy loss, making it suitable for use on edge devices (e.g., smartphones, tablets). Quantization-aware training incorporates quantization during training, producing models more robust to precision reduction. Courbariaux et al. (2015) introduced BinaryConnect, a quantization-aware method that trains neural networks with binary weights, significantly reducing memory usage and computational complexity.

Quantization methods are essential for complex tasks such as emotion-conditioned text generation, which enhances human-computer interaction but demands significant computational resources. Most research focuses on sentiment polarity (Zhang et al., 2019; Maqsud, 2015; Niu and Bansal, 2018), with fewer studies on emotion conditioning (Ghosh et al., 2017; Song et al., 2019; Zhou et al., 2018; Cui et al., 2022; Colombo et al., 2019; Menchaca Resendiz and Klinger, 2023).

None of the previous works have focused on the trade-off between affective text generation and quantization. Our research addresses this gap by investigating the trade-offs between GPU RAM, inference time, and text quality when using different precision bits in affective text generation.

2 Methodology

The presented work aims to provide an understanding of affective text generation using multiple precision bits[1]. Section 2.1 introduces the seed prompts use as as affective generation prompts for the LLMs. Section 2.2 describes the LLMs and the quantization configurations. Section 2.4 explains the automatic evaluation setup. Finally, Section 2.5 presents the text quality evaluation.

2.1 Seed Prompts

We select a set of 10 seed prompts strategically designed to guide the language models for effective affective text generation. This collection comprises a fusion of straightforward prompts (e.g., Generate a text that expresses `Emotion`, Phrases that express `Emotion`) that users can easily conceive, alongside prompts drawn from psychological studies, such as the ISEAR survey (e.g., "Please describe a situation or event—in as much detail as possible—in which a reader felt `Emotion`?"). The following diverse combination ensures a comprehensive exploration of emotional expressions, encompassing both intuitive and scientifically validated inputs.

1. Describe a situation where a person felt `Emotion`.
2. Write a text that expresses `Emotion`.
3. Phrases that express `Emotion`.
4. What is a sentence example for `Emotion`?
5. Can you provide an example of a situation where someone experienced `Emotion`?
6. What is an example of a `Emotion` sentence?
7. `Emotion` sentence.
8. Experience for `Emotion`?
9. Please describe a situation or event—in as much detail as possible—in which a reader felt `Emotion`.
10. Please complete the sentence: I felt `Emotion` when/because.

2.2 Precision Bits

To address the research question, **What is the trade-off between precision in bits on LLMs in affective text generation?**, we utilize LLAMA 2 Chat (Touvron et al., 2023) with three parameter sizes (7B, 13B, and 70B) and Mistral 7B (Jiang et al., 2023) and Mixtral-8x7B (Jiang et al., 2024). Each model is loaded into memory using three different precisions: 8, 16, and 32 bits – the default setting. We employ a total of 15 configurations to generate affective text from the seed prompts.

2.3 Text Generation

The text is generated using the LLMs and the 10 seed prompts. We design an *Emotion-Prompt* that integrates the seed prompt (for instance, "Write a text that expresses `Emotion`.") with one of seven specified emotions: Joy, Fear, Anger, Sadness, Disgust, Shame, and Guilt. In this setup, `Emotion` is replaced by each emotion category (see Table 1 for examples). For every *Emotion-Prompt*, we employ a beam search strategy to generate the three most likely sentences. This process uses a beam size of 30, sets the next-token temperature to 0.7, and applies a top-p (nucleus) sampling strategy at 0.7. Additionally, we take measures to ensure the generated output is free from sentences that contain duplicate instances of any bigram. The inference tasks for the LLMs models are performed and evenly distributed across a GPU cluster equipped with 8 NVIDIA RTX A6000 GPUs, for 10 times per model configuration.

[1]The code can be found at Blinded for review.

Emotion	Emotion-Prompt Example
Joy	Phrases that express `Joy`
Fear	Phrases that express `Fear`
Anger	Phrases that express `Anger`
Sadness	Phrases that express `Sadness`
Disgust	Phrases that express `Disgust`
Shame	Phrases that express `Shame`
Guilt	Phrases that express `Guilt`

Table 1: Examples of Emotion-Prompts for each emotion category for the prompt "*Phrases that express* `Emotion`.".

Emotion	Precision	Recall	F_1
Anger	.72	.58	.64
Disgust	.74	.80	.77
Fear	.78	.93	.85
Guilt	.56	.71	.62
Joy	.91	.92	.98
Sadness	.91	.87	.89
Shame	.66	.43	.52
Macro-Avg.	.75	.75	.75

Table 2: Precision, Recall and F_1 scores from the emotion classifier over the 7 emotion classes.

2.4 Automatic Classification

To evaluate the performance of the LLMs configuration automatically, we use an emotion classifier fine-tuned on the ISEAR dataset (Scherer and Wallbott, 1994). The emotion classifier is built on top of RoBERTa (Liu et al., 2019) using the standard parameters for ten epochs with a batch size of five. Please refer to Table 2 for precision, recall, and F_1 scores. Despite the limited data, these scores are comparable to previous experiments (Troiano et al., 2023).

2.5 Text Quality

We use GPT-3.5, known for its human-like performance in text quality evaluation (Chiang and Lee, 2023; Liu et al., 2023). This evaluation focuses on Coherence, Fluency, Grammar, Native Speaker Likeness, Human Likeness, and Plausibility. To quantify the evaluation, we adopt a 5-point Likert scale, ranging from 1 (disagree) to 5 (agree), to rate each dimension of text quality (Table 4).

3 Experimental Results

We investigate the trade-offs involve in using quantization with Large Language Models (LLMs) for affective text generation. Our evaluation includes the trade-offs between quantization and performance (Section 3.1), quantization and text quality (Section 3.2), performance and GPU (Section 3.3), and GPU and inference time (Section 3.4).

3.1 Quantization vs. F1-score in Affective Text Generation

The analysis between quantization levels and F_1-scores across different models shows multiple trends (see Table 3). Llama-2-7b's F_1-score drops at the highest quantization level (32), while Llama-2-13b shows variability but generally decreases at mid-level quantization (16). In contrast, Llama-2-70b's performance improves with higher quantization, suggesting a more efficient GPU resource usage. Mistral-7B maintains stable high F_1-scores across all quantization levels, showing robustness to changes in quantization. Mixtral-8x7B experiences a decline in performance with higher quantization, indicating sensitivity to reduce precision. The analysis suggest that the impact of quantization on LLM's performance varies, with some models benefiting from higher quantization levels and others experiencing a decline, guiding the choice of quantization based on desired performance and resource efficiency.

3.2 Quantization vs. Text Quality

Table 4 illustrates the trade-off between quantization and text quality. Quantization affects text quality differently across models. Llama models generally show decreased quality with higher quantization, especially in coherency and plausibility. Mistral-7b remains robust across all quantization levels, while Mixtral-8X7b improves with higher quantization. Larger models at lower quantization values typically outperform smaller models with higher precision – for example, Llama-2-70b at 8 bits outperforms Llama-2-7b or Llama-2-13b at 16 or 32 bits – while requiring similar memory.

Model	**Qnt.**	F_1	**GPU**	**I. Time** (std)
Llama-2-7b	8	.73	7.5	20.89 (.22)
Llama-2-7b	16	.73	13.4	12.53 (.82)
Llama-2-7b	32	.63	26.5	15.82 (1.8)
Llama-2-13b	8	.73	10.5	22.92 (.73)
Llama-2-13b	16	.65	25.6	11.54 (.08)
Llama-2-13b	32	.70	42.2	16.49 (.22)
Llama-2-70b	8	.64	50.4	34.93 (.97)
Llama-2-70b	16	.72	100.5	29.72 (.49)
Llama-2-70b	32	.74	268.1	51.93 (.28)
Mistral-7B	8	.73	8.9	69.99 (22.73)
Mistral-7B	16	.74	15.6	11.90 (.06)
Mistral-7B	32	.74	29.7	33.17 (.20)
Mixtral-8x7B	8	.72	46.5	125.04 (34.03)
Mixtral-8x7B	16	.66	90.9	85.89 (5.94)
Mixtral-8x7B	32	.66	181.2	53.56 (.53)

Table 3: Performance of Mistral and Llama models at different quantization (Qnt.) values. F_1-scores are computed across 7 emotions. GPU usage is in GB. Inference time (I. Time) is in seconds per generated text instance.

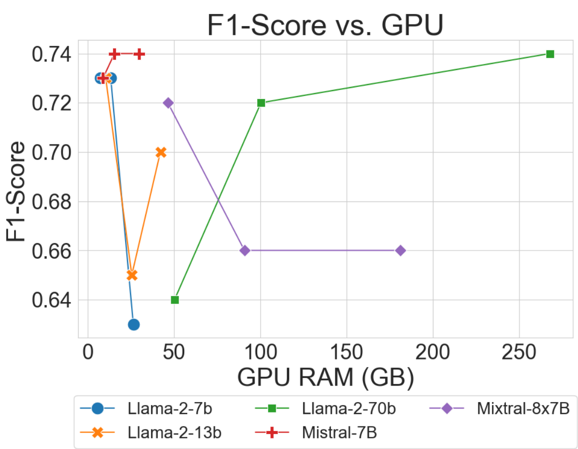

Figure 1: F_1-score vs. GPU memory usage for different models and configurations.

3.3 GPU Usage vs. F1-score

Figure 1 compares the performance of Llama and Mistral models across quantization levels (8, 16, and 32 bits) in terms of F_1 score and GPU RAM usage. Models with fewer parameters (Llama-2-7b and Mistral-7B) exhibit consistent F_1 scores across different quantization values. In contrast, models with more parameters (Llama-2-70b and Mixtral-8x7B) show greater variability in F_1 scores with changes in quantization. Generally, reducing model precision can decrease performance by up

Model	Qnt.	Flu	N. Spk.	Coh.	Pla.	Gra.
Llama-2-7b	8	3.3	3.6	2.4	3.0	3.5
Llama-2-7b	16	3.2	3.5	2.5	2.7	3.5
Llama-2-7b	32	3.2	3.5	2.2	2.4	3.5
Llama-2-13b	8	3.5	3.5	2.5	2.6	3.5
Llama-2-13b	16	3.1	3.6	2.3	2.5	3.5
Llama-2-13b	32	3.3	3.7	2.4	2.7	3.5
Llama-2-70b	8	3.7	4.0	3.2	3.5	3.8
Llama-2-70b	16	3.8	4.0	3.5	3.6	3.9
Llama-2-70b	32	3.4	3.6	2.4	2.8	3.6
Mistral-7b	8	4.1	4.0	3.8	3.8	3.8
Mistral-7b	16	4.1	4.1	3.9	3.8	3.9
Mistral-7b	32	4.1	4.1	4.0	3.7	3.9
Mixtral-8X7b	8	3.5	3.8	2.4	3.3	3.5
Mixtral-8X7b	16	4.1	4.0	4.0	3.7	3.9
Mixtral-8X7b	32	4.1	4.0	3.9	3.7	3.9

Table 4: Text quality evaluation using a five-level Likert scale for fluency (Flu), native speaker (N. Spk), coherency (Coh.), plausibility (Pla.), and overall grammar (Gra.). A score of 1 indicates "not at all agree," and 5 indicates "extremely agree" (higher is better).

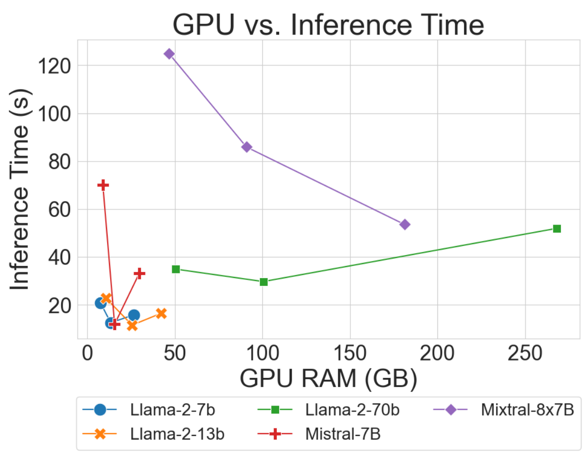

Figure 2: GPU usage (GB) vs. inference time (seconds) per text instance.

to 10 percentage points (Llama-2-70b) or increase it by 10 percentage points (Llama-2-7b), while reducing memory usage by approximately 76%. This indicates that in many scenarios, similar performance can be achieved with significantly lower GPU RAM usage.

3.4 GPU Usage vs. Inference Time

Table 3 and Figure 2 illustrate the comparative analysis of GPU usage and inference time for each model configuration. Generally, as quantization increases, inference time also increases, except for Mixtral-8X7b. More aggressive quantization leads to longer generation times. In most cases, generating an instance with 8-bit quantization takes about twice as long as with 32-bit quantization. However, this increased time is offset by a roughly 76% reduction in GPU RAM.

3.5 Qualitative Analysis

The comparative analysis (Appendix A, Table 5) reveals a qualitative distinctions in the text produced by the LLMs. While both the 7B and 70B models demonstrate similar performance in raw emotion scoring from the classifier, a closer examination of the output text suggests that the 70B models yield superior and more coherent results. This is evident in the more nuanced and detailed responses to prompts about emotions such as anger. The 70B-8 and 70B-32 configurations, in particular, generate text that more effectively captures the complexity of the emotional scenarios, providing richer context and depth in the examples given. Similar results can be observed with Mixtral-8X7b. For Mistral-7b, quantization has minimal impact on the generated text, with consistency maintained across different quantization levels, as show in previous sections. This suggests that larger models may have a better capacity for generating text that not only aligns with the desired emotion but also delivers a clearer and more precise narrative, indicative of a deeper understanding and processing of affective content.

4 Conclusion and Future Work

In summary, we explore the effects of quantization on LLMs for affective text generation.Our findings indicate that reducing precision bits leads to a performance decrease of up to 10 pp in F_1 score for larger models. In contrast, smaller models have an increase of up to 10 pp. While this reduction in precision does indeed affect model performance in affective text generation, it offers a substantial benefit in terms of memory consumption, with a remarkable reduction of 76%. In terms of text quality, larger models at lower quantization levels generally outperform smaller, higher-precision models while requiring similar memory. This trade-off between efficiency and efficacy is a critical consideration for deploying LLMs in resource-constrained environments where memory usage is a limiting factor.

Further research will explore the quantization trade-off in additional NLP tasks (e.g., prompt optimization), different languages, and quantization techniques (e.g., quantization-aware training).

Acknowledgements

Blinded for review.

5 Ethical Considerations

When deploying large language models (LLMs) for affective text generation, ethical considerations must be addressed, particularly in the context of quantization and its impact on model performance and resource usage. The reduction in precision and corresponding decrease in F_1 score for larger models may lead to less accurate or coherent text generation, which could affect user experience and trust, especially in applications requiring high sensitivity, such as mental health support or personalized recommendations. Additionally, the increased inference time associated with quantization could lead to slower response times, potentially frustrating users and reducing the effectiveness of real-time applications.

Finally, the underlying risks associated with the base pre-trained language models (e.g., Llama-2, Mistral) must be considered. These models may have been trained on biased data, potentially leading to text that perpetuates stereotypes or marginalizes certain groups. It's important to note that such risks are not inherent to the quantization methodology but stem from the base models used.

6 Limitations

While our study provides valuable insights into the effects of quantization on large language models for affective text generation, several limitations should be acknowledged. Our evaluation is limited to specific models (Llama-2 and Mistral), and results may vary with other architectures and parameter sizes. The study uses a single emotion dataset, and larger, more diverse datasets could offer more comprehensive insights. Additionally, while we noted an increase in inference time, the study does not deeply investigate computational overhead, potential optimizations, or the impact of different GPU hardware.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Cheng-Han Chiang and Hung-yi Lee. 2023. A closer look into using large language models for automatic evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8928–8942, Singapore. Association for Computational Linguistics.

Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 2019. Affect-driven dialog generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3734–3743, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David. 2015. Binaryconnect: Training deep neural networks with binary weights during propagations. In *Neural Information Processing Systems*.

Fuwei Cui, Hui Di, Lei Shen, Kazushige Ouchi, Ze Liu, and Jinan Xu. 2022. Modeling semantic and emotional relationship in multi-turn emotional conversations using multi-task learning. *Applied Intelligence*, 52(4):4663–4673.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2017. Affect-LM: A neural language model for customizable affective text generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 634–642, Vancouver, Canada. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. 2018. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2704–2713.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Umar Maqsud. 2015. Synthetic text generation for sentiment analysis. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 156–161, Lisboa, Portugal. Association for Computational Linguistics.

Yarik Menchaca Resendiz and Roman Klinger. 2023. Affective natural language generation of event descriptions through fine-grained appraisal conditions. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 375–387, Prague, Czechia. Association for Computational Linguistics.

Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, 6:373–389.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *OpenAI blog*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Klaus R. Scherer and Harald G. Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2):310.

Zhenqiao Song, Xiaoqing Zheng, Lu Liu, Mu Xu, and Xuanjing Huang. 2019. Generating responses with a specific emotion in dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3685–3695, Florence, Italy. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv e-prints*, pages arXiv–2307.

Enrica Troiano, Laura Oberländer, and Roman Klinger. 2023. Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Computational Linguistics*, 49(1).

Rui Zhang, Zhenyu Wang, Kai Yin, and Zhenhua Huang. 2019. Emotional text generation based on cross-domain sentiment transfer. *IEEE Access*, 7:100081–100089.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18, page 730–738. AAAI Press.

A Affective Text Generation

We generate affective text from five LLMs using three different configurations and the *Emotion-Prompt*. A beam search strategy was employ to produce the three most likely sentences, using a beam size of 30, a next-token temperature of 0.7, and a top-p (nucleus) sampling strategy set to 0.7. To ensure quality, we excluded sentences containing duplicate bigrams. Table 5 shows a sample of the generated text from different configurations using the same *Emotion-Prompt*.

Model	**Generated text**
Llama-2-7B-8b	"Here are some examples of situations that can cause anger. How to write a descriptive essay about a place. A descriptive essay is a type of writing that uses sensory details to describe a place, person, object, or experience. The goal of a descriptive essay is to create a vivid image in the reader's mind and to convey the writer's feelings and perceptions about the subject."
Llama-2-7B-32b	", but was unable to express it in a healthy way. A person may feel anger but be unable to express it in a healthy way due to various reasons. Here are some possible scenarios: 1. Fear of reprisal: The person may be afraid of the consequences of expressing their anger, such as losing their job, damaging their relationship, or facing social ridicule."
Llama-2-13B-8b	4. Describe a situation where a person felt fear. 5. Describe a situation where a person felt happiness. 6. Describe a situation where a person felt sadness. 7. Describe a situation where a person felt surprise. 8. Describe a situation where a person felt embarrassment. 9. Describe a situation where a person felt disgust.
Llama-2-13B-32b	What was the situation? What did they do with the anger? Did it change the outcome of the situation? 5. Describe a time when you were happy. What happened? Why were you happy? Did this happiness change your life in some way? 6. Describe a time when you were sad. What happened? Why were you sad? Did this sadness change your life in some way? 7. Describe a time
Llama-2-70B-8b	" * What was the trigger for the anger? * How did the person express their anger? * How did the situation unfold, and what was the outcome? * How did the person feel after the situation had resolved? * What did the person learn from the experience? Example: Situation: A colleague took credit for a project that Jane had worked on for months, and"
Llama-2-70B-32b	"You have been betrayed by someone you considered a friend. You trusted this person with your deepest secrets and they have gone behind your back and shared them with others. You feel angry and hurt by their actions."
Mistral-7B-8b	and Question 2: How did the person express their anger?
Mistral-7B-32b	and Question 2: How did the person express their anger?
Mixtral-8x7B-8b	and (a) expressed it openly, (b) suppressed it. Answer: (a) A student in my class was constantly disturbing the teacher during her lecture by talking loudly with his friends. The teacher asked him to keep quiet several times but he did not pay any attention. Finally, she lost her temper and scolded him in front of the whole class. She told him that if he didn't want to study, he should leave the class immediately. This made the student feel embarrassed and angry at the same time. However, he controlled his anger and apologized to the teacher for his behavior. (b) My younger brother had borrowed my new bicycle without asking me and damaged its chain while riding it roughly on the road. When I came to know about this, I felt very angry as the bicycle was quite expensive and I had saved money from my pocket money to buy it. However, I suppressed my anger as I didn't want to create a scene in front of my parents. I calmly explained to him how careless he had been and made him promise that he would never do such a thing again. I also made him pay for the repair of the bicycle from his own pocket money.
Mixtral-8x7B-32b	and a) The person was angry because they were not given the opportunity to speak in a meeting. They had prepared some important points that they wanted to share with the group, but the chairperson did not call on them. This made them feel frustrated and unheard, which led to feelings of anger. b) A friend of mine was angry because they found out that their partner had been cheating on them. They felt betrayed and hurt by this revelation, which caused them to experience intense anger towards their partner. c) I once felt angry when I was stuck in traffic for hours due to an accident on the highway. I had planned to arrive at my destination at a certain time, but the delay caused me to miss an important appointment. This made me feel frustrated and upset, leading to feelings of anger.

Table 5: Affective text examples generated by five LLMs at two quantization levels (8 and 32 bits) using the *Emotion-Prompt*: "Describe a situation where a person felt anger"