LLM Agent Controlling AI-Generated Images

A. Tibrewal, J. Kim, H. B. Goodman, M. Gupta, and V. Mittal

***Abstract*—In the flourishing and every growing scope of AI, text-to-image models hold immense potential, yet their success is hindered by the pivotal element of the quality of the input they receive. Current approaches often stumble due to inefficient, ambiguous prompts, leading to inaccurate outputs. We propose a novel framework that leverages the agility of small language models (SLMs) to overcome this bottleneck. By dynamically refining prompts based on user intent and context gleaned from curated examples, the SLM guides a diffusion model towards efficient and precise image creation. This paper explores methods for generating high-quality examples to fuel the SLM's learning and guide the diffusion model towards precise image creation.**

***Index Terms*—Large language models, small language models, in-context learning, prompt engineering, high-quality examples**

I. Introduction

GENERATIVE artificial intelligence (AI) has garnered significant attention as a potentially transformative technology [7]–[9]. Its primary objective is the creation of some output that adheres to a specified set of input criteria. The resultant output takes diverse forms, including natural language text (e.g., ChatGPT, Bard, Bing Chat, Claude), source code snippets (e.g., Copilot, Ghostwriter, Tabnine), music (e.g., MuseNet, MusicLM), and images (e.g., DALL-E, Midjourney, Stable Diffusion).

II. Related Works

Many generative AI systems rely on an input prompt—a natural language description outlining the objectives of the generated output [10], [11]. While more recent systems have embraced multi-modal inputs, incorporating elements like images in addition to prompts, the textual prompt remains the fundamental input mechanism. In such systems, supplementary inputs may furnish extra information or offer examples to guide the generation process. Nonetheless, the text prompt retains its central role in shaping output.

One limitation of generative AI systems lies in the fact that the output quality is contingent upon the formulation of the input prompt [12]–[15]. The AI community has developed various strategies, best practices, guidelines, and tools to assist in creating effective prompts. This process, known as prompt engineering, involves an iterative cycle of crafting, assessing, and refining prompts to achieve optimal results.

It's important to understand that prompt engineering is a task that varies significantly depending on the platform, presenting a significant challenge in the field [5]. The effectiveness of prompts can vary widely across different systems, with some prompts performing well in one environment but showing less optimal results in others [16], [17].

To tackle this challenge, we proposed the creation of a small language model (SLM). This model will focus on refining prompts through the application of In-Context Learning (ICL) techniques and utilize parameter-efficient prompt tuning methods. This approach aims to enhance prompt engineering and optimization.

Our proposed approach centers on the direct manipulation of the input prompt through the selection of a small language model. Notably, [1] highlights the remarkable achievements of small language models in performing analogous tasks while emphasizing their environmentally friendly nature. The study showcases that similar performance levels can be attained by adopting the methods elucidated in their paper.

To elevate our results further, we introduce high-quality examples, which we manually curate, serving as in-context learning demonstrations for the large language model [4]. This innovative approach streamlines the generation of additional examples, capitalizing on the data extracted from these demonstrations to fine-tune our small language model. In-context learning represents a pivotal methodology through which models can grasp concepts through analogical reasoning [2].

Following the implementation of in-context learning, our secondary baseline involves fine-tuning our small language model via parameter-efficient prompt tuning [6], a method that has garnered empirical support in previous research [3].

III. Background Knowledge

A. Stable Diffusion

Stable diffusion model is a generative artificial intelligence model that can be used to generate images from text. In our project, the prompts that we are supposed to optimize on will be passed as text into the stable diffusion model, which will in turn generate an image from the provided text.

B. Small Language Models

Small language models are typically scaled-down versions of large language models (LLMs). They have fewer parameters, but they share many of the same fundamental characteristics as LLMs. They are designed to be more computationally efficient and resource-friendly. The reduction in the number of parameters makes them more suitable for deployment on

This paper was submitted for review on December 17, 2023 to the instructors of the course CMSC421 at the University of Maryland, College Park.

A. Tibrewal is a student at the University of Maryland, College Park. (e-mail: atibrew1@terpmail.umd.edu).

J. Kim is a student at the University of Maryland, College Park. (e-mail: jkim1127@terpmail.umd.edu).

H. B. Goodman is a student at the University of Maryland, College Park. (e-mail: hbalickg@umd.edu).

M. Gupta is a student at the University of Maryland, College Park. (e-mail: mgupta25@terpmail.umd.edu).

V. Mittal is a student at the University of Maryland, College Park. (e-mail: vmittal@umd.edu).

resource-constrained devices and environments. The trade-off for the smaller model size is that small language models may not perform as well as LLMs on certain tasks, especially those that require extensive context or broad linguistic knowledge. However, they are often sufficient for many practical applications.

C. Large Language Models

These models are designed to understand and generate human-like text, making them valuable for a wide range of natural language processing tasks. They have the capability to predict the likelihood of word sequences or generate text based on a given input.

D. In-Context Learning

In-context learning refers to the ability of a model to adapt and make decisions based on the immediate context or user inputs. It enables the model to provide more relevant and context-aware responses, which is particularly important in natural language understanding and generation.

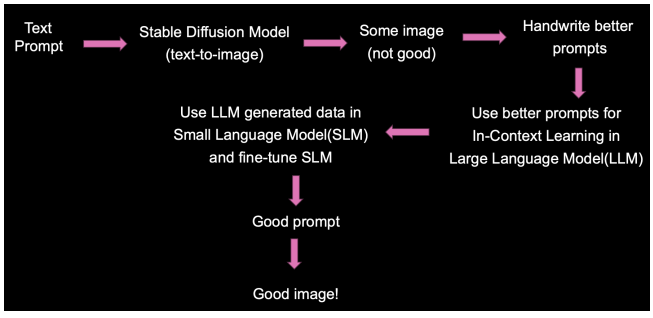
E. Parameter-efficient tuning

Parameter-efficient tuning is a set of techniques and practices designed to optimize the utilization of parameters in a model. The goal is to achieve good performance with a smaller number of parameters, making the model more computationally efficient, faster to train, and suitable for resource-constrained environments.

IV. Approach

The primary focus of this paper was the generation of high-quality examples through effective prompts. Our initial baseline strategy focused on implementing the diffusion model and generating exemplary instances. These were specifically designed to aid In-Context Learning (ICL), which was a crucial step in training the SLM. The second baseline aimed to leverage the ICL examples to construct a dataset encompassing both effective and ineffective prompts from a Language Model (LLM). Subsequently, our goal was to identify, implement, and train an SLM to execute prompt optimization. Additional objectives included prompt tuning and the application of necessary methods to further refine the performance of the SLM.

A complete flowchart schematic of our proposed approach is shown below:



In the early stages of our research, being newcomers to the field, we invested time in familiarizing ourselves with various terminologies discussed in the related works section below. Subsequently, we chose a pre-trained Stable Diffusion model and initiated the input of prompts from our selected training data—specifically, a Pokemon dataset featuring a diverse collection of cartoons and Pokemon characters, along with a starter prompt designed to generate their images. We have successfully developed five exemplary cases for In-Context Learning (ICL). Regrettably, due to pressing time constraints, we were unable to initiate our work on the second baseline.

V. Evaluation

A. What we evaluated

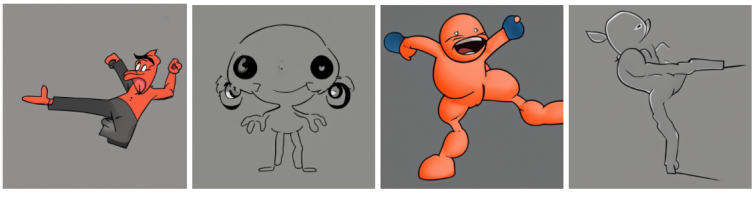
We assessed the effectiveness of the manually crafted prompts by measuring their proximity to the actual pictures corresponding to the prompts in our dataset. Below, we provide examples featuring original prompts, target images, transitions of model-generated images, and our handwritten prompts.

The subsequent illustrations of the side-by-side images depict the results obtinaed from tweaking the prompts, aiming to produce an image closest to the target. The last image in the transition series reflects our most successful attempt at achieving proximity to the target image.

B. Results

1) Example 1: Original Prompt: a drawing of a cartoon character doing a kick
Target Image:

Images obtained:



Optimized prompt: Create a drawing of a cartoon character doing a kick, similar to a Pokemon character performing a kick attack. The character should be gray in color and should be depicted in a dynamic and energetic pose, with their leg extended and their body leaning forward.

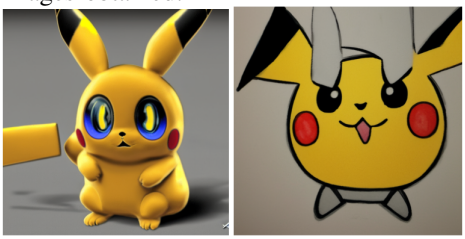
2) Example 2: Original Prompt: a drawing of a gray and black dragon
Target Image:

Images obtained:

Optimized prompt: a full-length drawing of a big gray and black dragon that looks like a rhinoceros. It has extremely small red eyes and small white claws and spikes on its back.

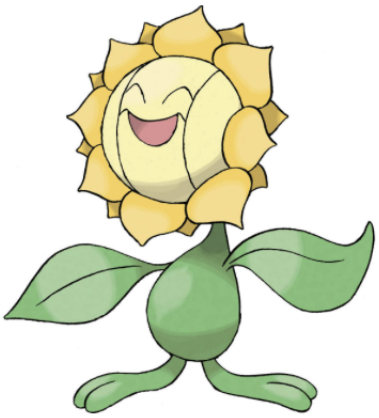
3) Example 3: Original Prompt: a cartoon pikachu with big eyes and big ears
Target Image:

Images obtained:



Optimized prompt: Generate a delightful cartoon rendition of Pikachu, emphasizing large expressive eyes and oversized ears. Capture the iconic charm of Pikachu in this illustration, ensuring that the character exudes a cute and endearing quality.

4) Example 4: Original Prompt: a cartoon sunflower with a happy face
Target Image:

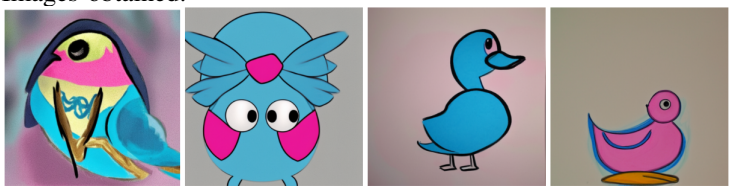


Images obtained:

Optimized prompt: A vibrant and cheerful cartoon illustration of a sunflower radiating positivity with a joyful expression. Accentuate happiness, ensuring the final result embodies the lively and playful spirit of a classic cartoon character. background should be plain

5) Example 5: Original Prompt: a pink and blue bird with big eyes
Target Image:

Images obtained:



Optimized prompt: Generate a cartoon drawing of a pink duck bird with blue tail and limbs and with prominent big eyes

VI. Conclusion

In conclusion, our team has successfully achieved the set objective in our initial baseline strategy. We effectively implemented the diffusion model and created a series of exemplary instances, each meticulously designed to support In-Context Learning (ICL). This approach has been instrumental in the training of the Small Language Model (SLM), marking a significant milestone in our project. Due to time conflicts and constraints, we only managed to complete our first baseline. The second baseline, where we train the SLM to refine prompts remains to be implemented, however, we have successfully completed the setup required to start the next baseline. Each member had an equal level of contribution towards documenting the research and this literature review. A link to our GitHub is posted in the Appendix section.

Acknowledgment

All the authors would like to thank Professor Tiyani Zhou and Hanyu Wang for their unwavering contribution and support.

References

[1] Schick, T., & Schutze, H., "It's not just size that matters: Small language models are also few-shot learners," *CoRR*, vol. abs/2009.07118, 2020. URL: https://arxiv.org/abs/2009.07118

[2] Dong, Q. et al., "A survey for in-context learning," *arXiv preprint arXiv:2301.00234*, 2022. URL: https://arxiv.org/abs/2301.00234

[3] Lester, B., Al-Rfou, R., & Constant, N., "The power of scale for parameter-efficient prompt tuning," *arXiv preprint arXiv:2104.08691*, 2021. URL: https://arxiv.org/abs/2104.08691

[4] Chan, S. et al., "Data distributional properties drive emergent in-context learning in transformers," *Advances in Neural Information Processing Systems*, vol. 35, pp. 18878–18891, 2022. URL: https://proceedings.neurips.cc/paper_files/paper/2022/hash/77c6ccacfd9962e2307fc64680fc5ace-Abstract-Conference.html

[5] Clarisó, R., & Cabot, J., "Model-Driven Prompt Engineering," *2023 ACM/IEEE 26th International Conference on Model Driven Engineering Languages and Systems (MODELS)*, pp. 47–54, 2023. URL: https://modeling-languages.com/wpcontent/uploads/2023/08/model-driven-prompt-engineering.pdf

[6] Oppenlaender, J., "A taxonomy of prompt modifiers for text-to-image generation," *arXiv preprint arXiv:2204.13988*, 2022. URL: https://doi.org/10.48550/arXiv.2204.13988

[7] Mullen, A. et al., "Top strategic technology trends for 2022: Generative AI," *Gartner Paper*, 2021.

[8] Wiles, J., "Beyond ChatGPT: The future of generative AI for enterprises," *Gartner Paper*, 2023. URL: https://www.gartner.com/en/articles/beyond-chatgpt-the-future-of-generative-ai-for-enterprises

[9] O'Grady, M., & Gualtieri, M., "Global AI software forecast," *Forrester report*, 2022.

[10] Brown, T. B., et al., "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[11] Liu, P., et al., "Pre- train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *Association for Computing Machinery*, vol. 55, no. 195, pp. 1-35, 2023. URL: https://doi.org/10.1145/3560815

[12] Zhao, W. X., et al., "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023. URL: https://arxiv.org/abs/2303.18223

[13] Ye, J., et al., "A comprehensive capability analysis of GPT-3 and GPT-3.5 series models," *arXiv preprint arXiv:2303.10420*, 2023. URL: https://arxiv.org/abs/2303.10420

[14] Reynolds, L., & McDonell, K., "Prompt programming for large language models: Beyond the few-shot paradigm," *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, no. 314, pp. 7, 2021. URL: https://dl.acm.org/doi/10.1145/3411763.3451760

[15] Witteveen, S., & Andrews, M., "Investigating prompt engineering in diffusion models," *arXiv preprint arXiv:2211.15462*, 2022. URL: https://arxiv.org/abs/2211.15462

[16] Petsiuk, V., et al., "Human evaluation of text-to-image models on a multi-task benchmark," *arXiv preprint arXiv:2211.12112*, 2022. URL: https://arxiv.org/abs/2211.12112

[17] Borji, A., "Generated faces in the wild: Quantitative comparison of Stable Diffusion, Midjourney and DALL-E 2," *arXiv preprint arXiv:2210.00586*, 2022. URL: https://arxiv.org/abs/2210.00586

Appendix

Github Link: https://github.com/hbalickgoodman/image-generation-CMSC421/tree/main

Presentation Link: https://docs.google.com/presentation/d/1c2_JstyIAkmDGS8taVxbL6Fh8g-90Jyz/edit?usp=sharing&ouid=103517508555274208113&rtpof=true&sd=true