Cognitive-Inclusive Communication: An Integrated Text and Graphics Generator for Improved Readability

Yong-Xiang Chang[a], Chiun-Li Chin[a], Zhong-Ting Zhemg[b], Geng-Kai Wong[b], Su-Juan Chen[a], Guan-Tsen Liu[a], Pei-Hsin Chang[c], Che-Cheng Liu[a], Pei-Chen Huang[a], Pei-Xin Ye[a]*
*e-mail: 1158032@live.csmu.edu.tw (P.X. Ye)

***Abstract*— People with intellectual disabilities frequently face challenges while reading due to their restricted reading abilities. Easy-to-read publications serve as valuable resources for this purpose. Therefore, this study explores the potential of LLaMA 2 13B transformer models to combine text with images to improve readability and expression. Also, we proposed a streamlined process that simplifies lengthy text using MLP layer to enable seamless integration with diffusion models. Our approach generates images and concise text that closely align with the original lengthy text. Through blind testing and evaluation, we conclude that the superiority of our method in accuracy and image quality is outdo another Stable Diffusion manner. This study underscores the significance of integrated models, dimensionality reduction, and end-to-end processing in enhancing information accessibility.**

Keywords —reading abilities, Easy-to-read, LLaMA 2 13B, readability, MLP layer, diffusion model, concise text, end-to-end.

I. INTRODUCTION

Individuals with intellectual disabilities often encounter difficulties while reading due to their limited reading abilities. For these persons, easy-to-read publications can be a door-opener and a useful training resource [1]. Meanwhile, the transformer model is mainly used in the fields of natural language processing and computer vision. Supposed that it can use multi-modal processing capabilities to combine text and images, thereby enhancing the readability and expressiveness of information. Following the Transformer framework, we choose LLaMA 2 13B as a suitable replacement for the closed-source model [2]. Since LLaMA 2 13B is superior to other large-scale language models, its open-source value and model adjustable features will make it suitable for combination with diffusion models that generate images from text [3]. And it proposed that by introducing MLP, the overall process will be optimized and followed end-to-end [4].

II. METHODS AND RESULTS

This study aims to achieve the simplification of lengthy text in an end-to-end manner. This process will include diminishing the dimensionality through MLP layer, resulting in a size appropriate for input into the diffusion model. The task is to present an image that best corresponds to the original text, along with a condensed concise text, to achieve an easy-to-read effect. The lengthy text as input and opt to utilize the LLaMA 2 13B model as our foundational model. The results of the last layer of the encoder are taken out and fed into the MLP layer. The core focus of MLP layer is for reducing the last attention normalization output of LLaMA 2 to 59136(77×768) text embeddings. This represents the closest feature vector to the concise text that matches the Context Embeddings of our diffusion model. The whole processing is illustrated by Figure 1.

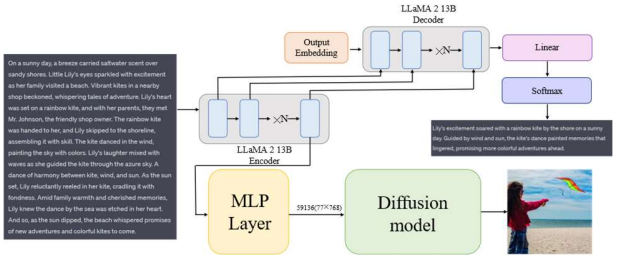

Figure 1. Output Context Embeddings Processing with MLP Layer.

There are two different ways to accomplish the same lengthy text to generate an image. The first approach is using our above subject, which achieves the end-to-end process. The second approach involves subjecting the lengthy text to two passes of the LLaMA 2 13B model to generate a prompt, which stems from the concise text produced during the first pass. This prompt is then fed into the stable diffusion model to generate corresponding images. It would submitted the images generated in both ways to multiple rounds of blind testing and assessment by image recognition experts. The evaluation process will focus on the degree of correspondence between the images generated by the two methods and the original concise text. The non-end-to-end way is much inferior matching due to probability weight distribution.

III. CONCLUSION

In this study, the primary objective to approach cognitive-inclusive communication combines transformer models and image generation to enhance accessibility for individuals with intellectual disabilities. Through streamlined processes and the integration of MLP layer, it achieves easy-to-read functions. Different from the existing non-end-to-end methods, it could connect the graph with concise text very well. Our proposed model superiority over the diffusion model in blind testing underscores its potential in generating images closely aligned with concise text.

REFERENCES

[1] M. Nomura, G. Skat Nielsen, and B. Tronbacke, "Guidelines for easy-to-read materials." *International Federation of Library Associations and Institutions (IFLA)*, No. 120, 2010.

[2] H. Touvron, et al. "Llama 2: Open foundation and fine-tuned chat models." *arXiv preprint arXiv:2307.09288*, 2023.

[3] R. Rombach, et al. "High-Resolution Image Synthesis with Latent Diffusion Models." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,* pp. 10684-10695, 2022.

[4] O. Ilya Tolstikhin, et al. "Mlp-mixer: An all-mlp architecture for vision." *Advances in neural information processing systems*, vol. 34, pp. 24261-24272, 2021.

a. Medical Informatics, CSMU, Taichung, Taiwan
b. Department of Automatic Control Engineering, FCU, Yaichung, Taiwan
c. XSSH, Taipei, Taiwan