

Multilingual Text Enhancer (MLTE) – A LLaMA2 based model for prompt generation

NV Sai Teja[#], Kuldeep Kumar, Muthukumaran Malarvel*

Department of Computer Science and Engineering
Hindustan Institute Of Technology and Science, Chennai, India

*Corresponding author email: mmuthu@hindustanuniv.ac.in

[#]Student Member IEEE ^{*}Senior Member IEEE

Abstract— This paper introduces MLTE - Multilingual Text Enhancer, a text enhancement model developed primarily to the enhance input text for text-to-image generation models. The existing text encoders in image generation models often have limited capabilities, and the quality of image generation depends on the prompt. If the prompt includes misspelled words, the encoder will create an irrelevant image. Most encoders are primarily based on English. MLTE effectively addresses multilingual prompts, misspelled words, overly verbose prompts, and creatively enhances the prompt to achieve improved results. MLTE employs sophisticated natural language processing algorithms to establish a link between unprocessed textual input and the production of highly precise visual content. MLTE is based on LLaMA2, which has the ability to handle numerous languages and enables the simple incorporation of content from diverse linguistic origins. Additionally, its spell-checking and correction functions ensure the quality and coherence of the prompt. Moreover, MLTE's scene and text augmentation features strengthen the visual richness and coherence of generated photos, thereby enhancing their overall quality and realism. Its summarizing capability condenses large paragraphs into concise yet helpful summaries, which assisting the image creation process by delivering more focused inputs. MLTE can be used with any text to image generating models. By undertaking empirical evaluation, this paper demonstrates the effectiveness of MLTE in boosting text quality for text-to-image synthesis tasks, leading to significantly improved image generation results.

Keywords— Text-to-Image Generation, Multilingual Text Processing, Natural Language Enhancement, Image Synthesis, LLaMA2, Text Summarization

I. INTRODUCTION

In recent years, Text-to-Image (T2I) generative models have witnessed significant advancements owing to both algorithmic innovations and the availability of extensive paired training datasets [1]. However, the efficacy of these models in generating high-quality images depends upon the precision and coherence of the textual prompt description. Presently, existing text encoders, such as CLIP [2], predominantly operate within the confines of the English language, thereby limiting their applicability to multilingual contexts. These encoders often exhibit constraints in vocabulary diversity and linguistic coverage. Moreover, inherent text encoders within text-to-image generative models face challenges in accommodating prompts containing misspelled words, non-English languages, insufficient prompt details, or excessively verbose prompts. To address this challenge, this paper introduces MLTE -

Multilingual Text Enhancer, based on LLaMA2 7B model. A text enhancement model tailored to refine and augment input text for text-to-image generation tasks. This paper primarily focuses to solving: 1. multilingual prompts (prompts in non-English languages), 2. misspelled words, 3. enhancing insufficient prompt details, 4. to summarize and keyword extraction for lengthy prompts. LLaMA2 is trained on less than 4.5% non-English datasets [3]. LLaMA2 uses byte-pair encoding (BPE) algorithm [14], its multilingual support enables MLTE to handle diverse linguistic inputs, broadening the scope of text-to-image generation across various cultural contexts. MLTE integrates robust spell-checking and correction mechanisms to enhance text accuracy and readability, thereby improving the overall description quality and image synthesis. We applied our MLTE model on Dalle-3, Midjourney, and Stable Diffusion image generation models to evaluate the quality of images and compare how well our MLTE model handled the multilingual prompts.

II. RELATED WORK

The landscape of text-to-image generation has been significantly influenced by various frameworks, with GANs (Generative Adversarial Nets) [13] emerging as a prominent methodology. Subsequently, advancements such as Stack-GAN [4], Attn-GAN [5], and SD-GAN [6] have demonstrated promising results in this domain. Recent research has further elevated image generation quality, with works like DF-GAN [7] introducing fusion modules for effectively integrating text and image features. Notably, LAFITE [8] leveraged CLIP's model to construct pseudo image-text pairs, thereby enhancing the efficiency of GAN-based text-to-image synthesis. In addition to advancements in GAN-based methods, diffusion models have also garnered attention, with approaches like GLIDE [9] demonstrating guided inference for high-fidelity image generation. DALL-E-3 [2] and Imagen [1] have set new benchmarks in text-to-image generation by employing diffusion models in conjunction with innovative techniques such as learning CLIP image embeddings and leveraging conditional diffusion models for mapping text embeddings to images. While most encoders in image generation models are primarily English-based, there is still significant room for improvement, particularly concerning the performance limitations of text encoders.

Large language models (LLMs) [3][15][16] have demonstrated performance in various fields such as text generation, coding, translation, summarization, and many other domains. LLMs like GPT, Gemini, Mistral, and LLaMA2 are hugely based on English. However, their performance in other languages are limited. To overcome this limitation, researchers have developed a new approach: Multilingual Large Language Models (MLLMs). These models can process text in multiple languages simultaneously by learning a unified representation during pre-training on massive multilingual datasets. This capability allows them to adapt to specific tasks or languages through fine-tuning. Representative works include mBERT [17], and the recent advancements such as BLOOM [18] and PolyLM [19] have made significant progress. However the models are very heavy. Another line of research adopted existing monolingual LLMs to multilingual using techniques such as prompt engineering [20].

III. METHODOLOGY

A. Structure of the project

This paper utilized the LLaMA2-7B parameter version of Meta's open-source LLM (Large Language Model) as the foundation of our model. This specific version of the model has been pre-trained on extensive datasets and is renowned for its proficiency in generating safe content at this size [3]. After incorporating the synthesized dataset, we proceeded to fine-tune the LLaMA2 model to generate prompts for Text-to-Image generation models. This approach initially involved performing Supervised Fine-Tuning with Quantized Low-Rank Adaptation (QLoRA - a Parameter-Efficient Fine-Tuning method) [11] on the base model followed by Reinforcement Learning techniques and further optimization through Instruction Fine-Tuning to achieve superior results.

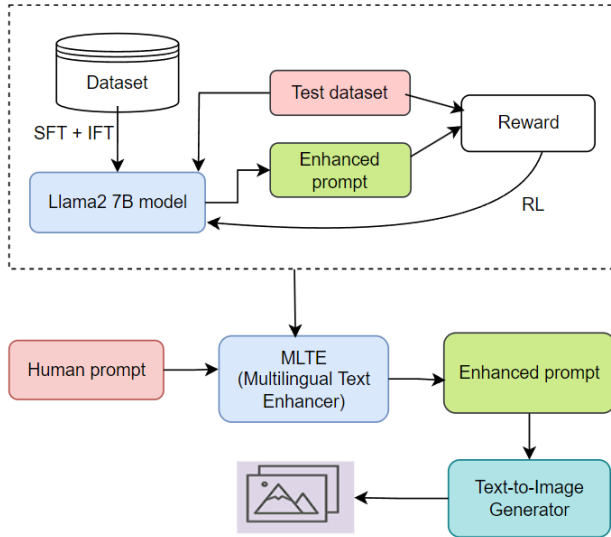


Fig. 1. Architecture Diagram

B. Dataset

This paper works with three datasets. The first one is **multilingual** dataset with 10k non-English to English translations, which are used to teach the vocabulary for the model. The second is the **multilingual_prompt**, comprising 1k prompts in two Indian languages, Telugu and Hindi. It consists of two columns: multilingual prompt and enhanced prompt. Since there was no existing multilingual prompt data available, we leveraged ChatGPT 4 to manually generate and verify prompts to create a synthetic dataset. The third dataset is the **Midjourney dataset**, containing 900k prompts across different styles and fields of image generation.

పసుపు పొగమంచుతో కూడిన పర్వత శ్రేణులు	A watercolor painting of a vibrant blue lake nestled amidst majestic mountains shrouded in yellow mist.
बर्फ से ढके पहाड़ों की चोटी पर स्थित	An ancient temple perched on the peak of snow-capped mountains, bathed in the golden rays of sunrise. A magnificent blend of pink and orange hues paints the sky.
ఆకాశంలో ఎగిరే రంగురంగుల వాతావరణం	A sea of colorful hot air balloons soaring through the sky, with lush green hills and flower-filled fields sprawling beneath.
एक सुनहरे रेतीले समुद्र तट पर एक	A golden conch shell resting on a pristine golden sand beach, glistening from the gentle touch of waves. In the background, coconut trees sway in the breeze with a captivating sunset view.

Fig. 2. Synthetic dataset (multilingual_prompt)

C. Fine Tuning

1) Pretrained LLaMA2 :

The LLaMA2 (Large Language Model Meta AI) is an autoregressive model based on the Transformer architecture [3]. It is trained on a massive dataset of approximately 2 trillion tokens encompassing diverse fields such as language, code, and mathematics. The data undergoes tokenization using the Byte Pair Encoding (BPE) algorithm [14], implemented through Sentence Piece [20]. BPE is a versatile tokenization method that can handle words from any language, eliminating the need for unknown tokens and separate number tokenization. It decomposes unknown UTF-8 characters into individual bytes and represents Unicode characters as one to four one-byte code units. This method prioritizes sharing vocabulary across languages, leading to improved performance. It goes a step further by enabling knowledge transfer between languages, even those with entirely different character systems.

2) SFT: Supervised Fine-Tuning :

We utilized Google Colab platform, equipped with T4 GPU boasting 16 GB of VRAM, to conduct fine-tuning on the LLaMA2 7B model, which contains over 7 billion parameters. To augment the model's comprehension of grammar and vocabulary in new languages, we employed a **multilingual** and **Midjourney** dataset. We fine-tuned the

model using QLoRA, which initially quantizes the pre-trained language model to 4 bits [11] and then locks those parameters. Additionally, the model is furnished with several trainable Low-Rank Adapter layers. The use of 4-bit quantization via QLoRA facilitates efficient fine-tuning of LLaMA2 while upholding high performance. Throughout the fine-tuning process, gradients are propagated backwards through the frozen 4-bit quantized model, affecting solely the Low-Rank Adapter layers [10]. Consequently, the entire pre-trained model remains fixed at 4 bits, with only the adapters being subject to updates.

Remarkably, the adoption of 4-bit quantization does not compromise model's performance, and QLoRA selectively updates a subset of parameters while effectively freezing the rest.

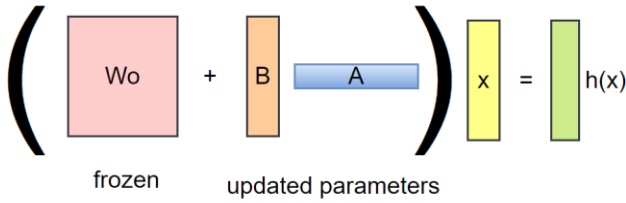


Fig. 3. QLoRA Rank matrix

$$h(x) = W_0 x + \Delta W x \quad (1)$$

In equation (1), $h(x)$ represents the hidden layer [11]. W_0 denotes the frozen parameters, while ΔW signifies the updated parameters, where $\Delta W = AB$. Here, A and B are low-rank matrices capturing the updated parameters. In this paper, a rank of 64 is utilized.

We utilized a cosine learning rate with an initial learning rate of 2×10^{-4} , training batch size of 4, a weight decay of 0.001, a rank as 64, and a sequence length of 4096 tokens for 1 epoch. During training, each sample combines a prompt x and an enhanced prompt y . To manage sequence length, we merged all user prompts and enhanced prompts from the training data. A specific template recommended by the developers of LLaMA2 is used to distinguish between the original and improved prompt sections. We leverage an autoregressive objective function, but we exclude the user prompt tokens from the loss calculation. This ensures that backpropagation focuses solely on the generated answer tokens.

3) Reinforcement Learning :

To train the reward model, we transform the collected human-generated data pairs into a format suitable for binary ranking. This involves converting them into labels such as 'chosen' (c) and 'rejected' (r), ensuring that the chosen response receives a higher score compared to its counterpart. To achieve this, we utilize a binary ranking loss function [12].

TABLE I. MODEL SCORES FOR DIFFERENT PTOMPTS

Model	User prompt /MLTE prompt	Relatively correct image				Unrelated image or No image			
		T1	T2	T3	T4	T1	T2	T3	T4
Stable diffusion	User prompt	0.08	0.64	0.88	0.76	0.92	0.36	0.12	0.24
	MLTE prompt	0.6	0.76	0.92	0.96	0.4	0.24	0.08	0.04
Dalle 3	User prompt	0.44	0.72	0.96	0.88	0.56	0.28	0.04	0.12
	MLTE prompt	0.68	0.88	0.92	0.92	0.32	0.12	0.08	0.08
Midjourney	User prompt	0.24	0.56	0.92	0.8	0.76	0.4	0.08	0.2
	MLTE prompt	0.55	0.92	0.92	0.92	0.45	0.08	0.08	0.08

$$L_{\text{ranking}} = -\log(\sigma(r\theta(x, y_c) - r\theta(x, y_r))) \quad (2)$$

In equation (2), $r\theta(x, y)$ [12] represents the scalar score output for prompt x and enhanced prompt y with model weights θ . y_c denotes the preferred response selected by annotators, while y_r represents its rejected counterpart. We utilized this score to conduct reinforcement learning on our model.

4) IFT: Instruction Fine-Tuning :

We utilized the **multilingual_prompt** dataset, which contains multi-language prompts and enhanced prompts pair. The dataset was formatted to adhere to the Llama2 instruction structure, which includes system instructions, user prompts, and enhanced prompts for each query. This approach resembles supervised fine-tuning but introduces an additional element: specific system instructions tailored to each query. These instructions guide the model towards a deeper understanding of the context and enable it to respond appropriately to user prompts across various scenarios. Finally, our model is fine-tuned for prompt enhancement, achieving a training loss of 3.2247

$$MLTE = LLaMA2 \text{ 7B} + SFT + RL + IFT$$

IV. RESULT AND ANALYSIS

The MLTE model can be applied to any Text-to-Image generation models for improved results. Therefore, we considered three different models (Stable Diffusion, Dalle 3, and Midjourney models), which are currently popular and efficient.

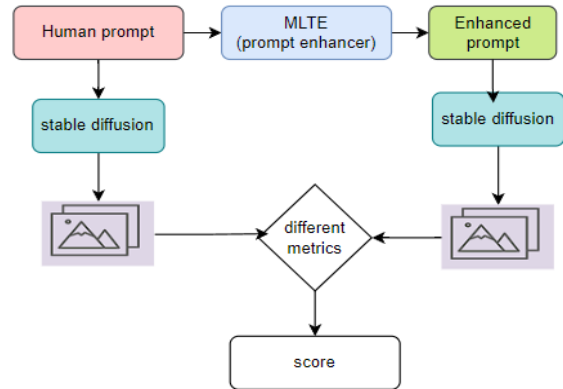


Fig. 4. Evaluation method

PROMPTS

User prompt:

दो बच्चे क्रिकेट खेल रहे हैं

MLTE prompt:

Two kids are playing cricket in the field

DALLE 3



Fig. 5(a)



Fig. 6(a)

MIDJOURNEY



Fig. 5(b)



Fig. 6(b)

STABLE DIFFUSION



Fig. 5(c)



Fig. 6(c)

Example of a T1 sentence: The user input is in Hindi, The T2I models did not encode words correctly, Fig. 5(a) is partially correct but missed **two kids**, Fig.5(b) and 5(c) are irrelevant.

User prompt:

dogg is bwarking

MLTE prompt:

A photorealistic image of an angry Golden Retriever mid-bark, with its mouth wide open and teeth bared



Fig. 7(a)



Fig. 7(b)



Fig. 7(c)



Fig. 8(a)



Fig. 8(b)



Fig. 8(c)

Example of a T2 sentence: The user prompt contains misspelled words. The models understood the misspelled word 'dogg' but missed the **barking**. (See Fig. 7(a),7(b),7(c).) The MLTE prompt captured both words 'dogg' and 'bark' and enhanced prompt with context, such as 'Golden Retriever.'

User prompt: (200 words)

A lonely elephant was looking for friends in the forest..... The elephant proceeded to kick the tiger and scare it away.

MLTE prompt :

A mighty elephant stands in a sun-dappled forest clearing, trumpeting and **kicking a frightened tiger** that flees. Relieved smaller animals, including a monkey, rabbit, and frog, watch from the background



Fig. 9(a)



Fig. 9(b)

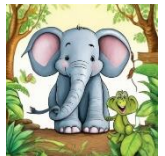


Fig. 9(c)



Fig. 10(a)



Fig. 10(b)



Fig. 10(c)

Here's an example of a T3 sentence: The user prompt consists of 200 words. However the models only processed the first sentence and ignored rest, resulting in images of **innocent elephant**. The MLTE prompt captures important keywords.

User prompt:

Polaroid photo of a young man wearing a suit.



Fig. 11(a)



Fig. 11(b)



Fig. 11(c)

MLTE prompt:

Polaroid photo of a young man wearing a suit and tie, holding a cup of coffee, in mid-dance move, Full body shot, in a bustling city square during sunset, Golden hour light casting a warm glowing



Fig. 12(a)



Fig. 12(b)



Fig. 12(c)

Example of a T4 sentence: The user prompt lacks context. MLTE prompt enriched the prompt by adding with related keywords such as **tie**, **coffee**, **building**.

The paper assesses image generation models with and without text enhancement across four prompt categories, yielding 600 images. Images are classified as relatively correct, unrelated, or not generated. A relatively correct image shares at least one keyword with the prompt; otherwise, it is deemed unrelated. The absence of images often results from unrecognizable non-English words triggering NSFW filters or exceeding token limits. The paper documents image generation frequencies for each text category with and without enhancement, highlighting the challenges encountered by AI art generators, as summarized in Table 1. Score = (number of images generated) / (total number of test prompts). For example, using Stable Diffusion model in category T1 using text enhancer if it generate 15 correct or related image then then $15/25=0.6$. For example, using Stable Diffusion model in category T1 using text enhancer, if it generates 15 correct or related image, then the score would be $15/25=0.6$

We ensured that the keywords in the test dataset for the T1 sentence are known to the model i.e., the prompts we used for testing contain at least one keyword present in the training dataset. For example, if the training dataset contains “red color ball”, we utilized “red color car” in the testing dataset.

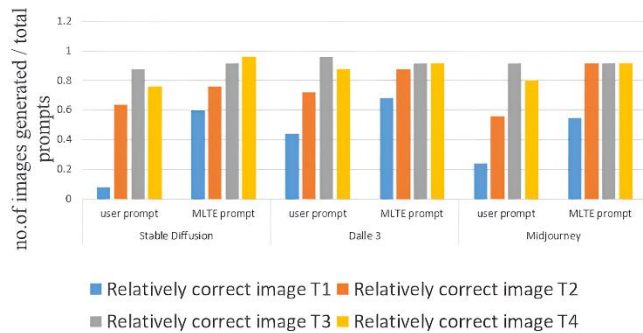


Fig. 13. Improvement in image generation using MLTE.

Fig. 13 graph indicates the effectiveness of text-to-image (T2I) models with and without Multilingual Text Enhancers (MLTE). While some T2I models, such as DALL-E 3 (Fig. 5(a)), can generate partially accurate images without MLTE (e.g., capturing a cricket scene but lacking two children for the prompt “two kids playing cricket”), the proposed MLTE approach consistently improved results across four prompt categories. The improvement in T4 is very little. The primary improvement was observed in the T1 prompt category, where the baseline model struggled. With MLTE, approximately 50% of generated images are relevant to the prompt. The main reason for remaining irrelevant images is that the prompt enhanced by MLTE is imperfect. As shown in Fig. 14, for some prompts, MLTE partially translated the text and it mixed with different languages. For some keywords, it is unable to translate into English due to lack of knowledge. For some sentences, if it does not know the English, it just kept the same original script keyword. This is due to imperfect encoding. This can be solved by training the model with a more diverse dataset and knowledge so that it can differentiate between languages.

Kshtigrast अंतरिक्ष यान ki khirki se ek akele अंतरिक्षयात्री ko
praayaanaaన్ని prakatించే oka
ప్రకటించే ఒక పాత ట్రావెల్ పోస్టర్

Fig. 14. MLTE model mix multiple languages

To measure the overall image quality, aesthetics, completeness, and correlation between the prompt and generated image, we conducted a human evaluation survey. we asked people which image they preferred for the given prompt. we considered only those images for which both prompts (user prompt and MLTE prompt) generated relatively correct images.

We used the same set of 200 prompts for all three image generation models, i.e., $(25 \times 4 = 200)$ the same 200 prompts for the three models. This method aided in evaluating and comparing the quality of image generation. We obtained

similar types of images and styles for comparison. We conducted a blind test by not disclosing the prompt type and image generation model to avoid bias, and we asked 15 human evaluators to vote for the better image for the given prompt.

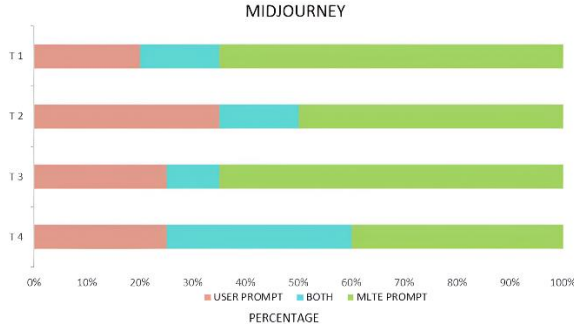


Fig. 15. Human preferred images for Midjourney

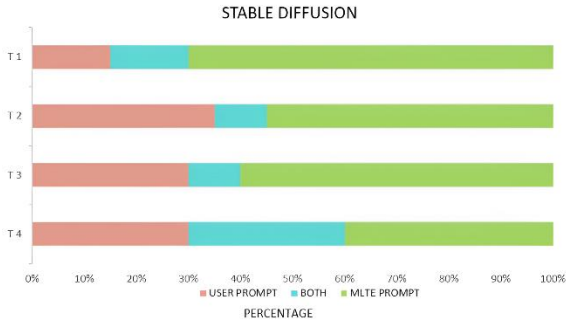


Fig. 16. Human preferred images for Stable Diffusion

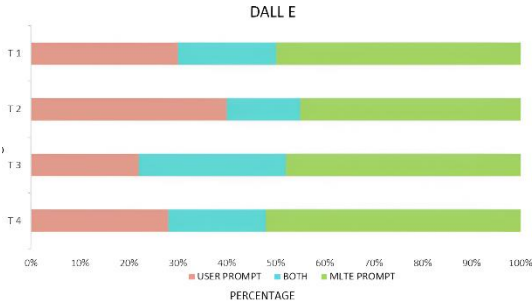


Fig. 17. Human preferred images for Dalle 3

Figures (15), (16), (17) has demonstrated that people generally prefer images generated by MLTE. MLTE has enhanced image generation quality across all prompt types with all three types of text to image generation models.

V. CONCLUSION

MLTE - Multilingual Text Enhancer has emerged as a solution to the limitations found in current text encoders used in text-to-image generation models. It tackles obstacles such as prompts in different languages, words with errors, and long inputs, thereby enhancing the quality of the resulting images. With the fine-tuning of LLaMA2, MLTE achieved a high level of coherence and accuracy in Telugu and Hindi linguistic contexts. Additionally, the spell-checking, summarization, and augmentation features greatly enhance the realism and depth of the generated images. The efficiency

of image generation was greatly improved by MLTE's capability to condense lengthy paragraphs into brief yet informative summaries. Empirical evaluations have highlighted the impressive capabilities of MLTE in improving text quality for image synthesis tasks.

REFERENCES

- [1] Chitwan Saharia, William Chan, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022.
- [2] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark . Hierarchical text-conditional image generation with clip latent, 2022.
- [3] Hugo Touvron, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models, 2023.
- [4] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In ICCV, 2017.
- [5] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In CVPR, 2018.
- [6] Guojun Yin, Bin Liu, Lu Sheng, Nenghai Yu, Xiaogang Wang, and Jing Shao. Semantics disentangling for text-to-image generation. In CVPR, 2019.
- [7] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. Df-gan: A simple and effective baseline for text-to-image synthesis. In CVPR, 2022.
- [8] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Lafite: Towards language-free training for text-to-image generation., 2022.
- [9] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741, 2021
- [10] Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen. Lora: Low-Rank Adaptation Of Large Language Models, 2021
- [11] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman Luke Zettlemoyer. QLORA: Efficient Finetuning of Quantized LLMs, 2023
- [12] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. In NeurIPS, 2020.
- [13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. NeurIPS, 2014.
- [14] Daeseung Park, Gi-taek An, Chayapol Kamyod, Cheong Ghil Kim, A Study on Performance Improvement of Prompt Engineering for Generative AI with a Large Language Model, 2024.
- [15] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways, 2022.
- [16] Tom Brown, Benjamin Mann, Amanda Askell, et al. 2020. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901.
- [17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.
- [18] Teven Le Scao, Angela Fan, Christopher Akiki et al. Bloom: A 176b parameter open-access multilingual language model, 2022.
- [19] Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li et al. PolyLM: An open source polyglot large language model, 2023.
- [20] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman et al. Crosslingual generalization through multitask finetuning, 2023.