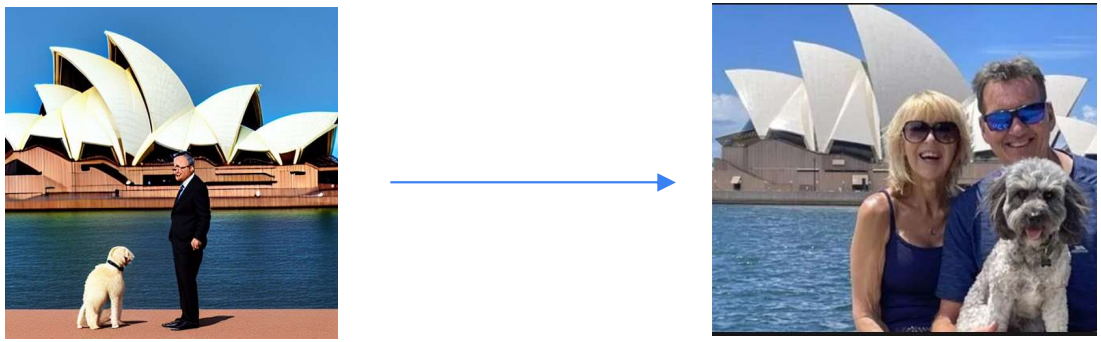
Improving Diffusion-based Image Generation Model with RLHF

Rongshang Li 500179600
MPhil.
Supervisor(s):Prof. Dacheng Tao

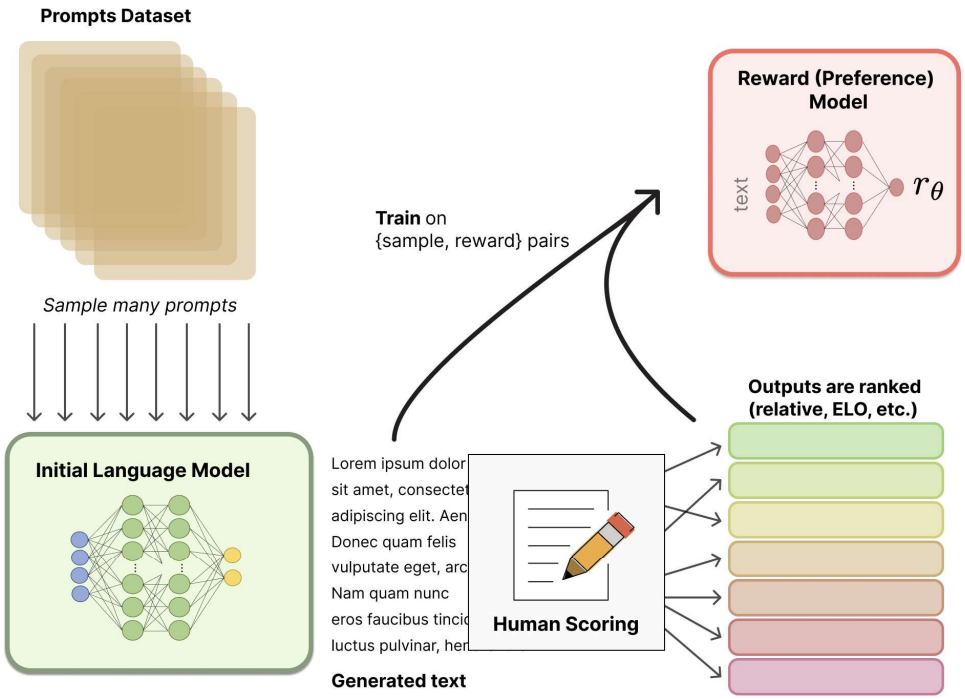
My Research

- Improving generation quality of pretrain generative models.
- Specifically: using RLHF(reinforcement learning from human feedback) to debias/ prevent harmful content/ improve consistency.



Motivation

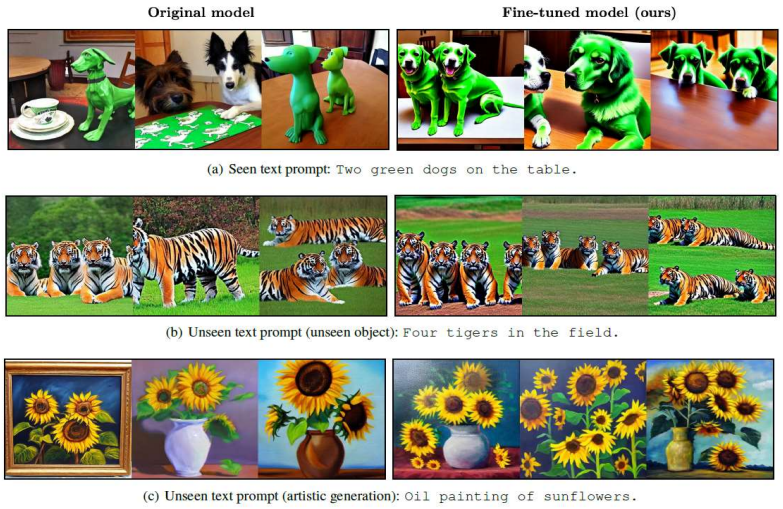
- RLHF works for ChatGPT
- But RLHF for ChatGPT needs a huge amount of human annotations.
- How to make the best use of the resources we have?



Literature Review

- 2 methods have been tried:

- Use pre-trained models to generate images, which are then finetuned based on user feedback.
- or
- Adopt real-time user feedback to update certain image



Research Method

Fine-tune the pre-trained model using feedback from RLHF-trained language models (i.e. ChatGPT).
Using API provided by ChatGPT to generate rankings of generated images.

Drawbacks

Is feedback from ChatGPT good enough to simulate “Human Feedback”?
Has OpenAI already completed it? (DALL-E 3)

Project Plan

Date	Tasks	Deliverables
Oct 2023-Nov 2023	Coding for Automatic Annotation system using ChatGPT API	A runnable system.
Nov 2023-Jan 2024	Testing the system and analyze the results.	Test results and analyzation.
Jan 2024-Mar 2024	Organize the results into a paper	Paper

Thanks!

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” 2022. 2
[2] X. Wu, K. Sun, F. Zhu, R. Zhao, and H. Li, “Human preference score: Better aligning text-to-image models with human preference,” 2023. 6, 7
[3] Z. Tang, D. Rybin, and T.-H. Chang, “Zeroth-order optimization meets human feedback: Provable learning via ranking oracles,” 2023. 6, 7
[4] K. Lee, H. Liu, M. Ryu, O. Watkins, Y. Du, C. Boutilier, P. Abbeel, M. Ghavamzadeh, and S. S. Gu, “Aligning text-to-image models using human feedback,” 2023. 6
[5] Y. Fan, O. Watkins, Y. Du, H. Liu, M. Ryu, C. Boutilier, P. Abbeel, M. Ghavamzadeh, K. Lee, and K. Lee, “Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models,” 2023. 6, 7
[6] https://openai.com/dall-e-3
[7] Lambert, et al., "Illustrating Reinforcement Learning from Human Feedback (RLHF)", Hugging Face Blog, 2022.