REENTREGA DE

TRABALHO DE FINAL DE SEMESTRE

Disciplina: Engenharia de Prompts
Professor: Thiago Ciodaro Xavier

Nome do Aluno: Matheus de Oliveira Stichtenoth

Caxias do Sul - RS
2024

Reentrega

Atividade 9:

Questão Letra A: Descreva o funcionamento dos modelos de imagem, segundo suas arquiteturas, limitações e vantagens:

Stable Diffusion:

Esse modelo de criação de imagens usa uma abordagem chamada de “difusão estável”, para replicar imagens de alta eficiência a partir dos prompts gerados pelo usuário.
Essa abordagem é baseada nos conceitos de difusão da física, onde partículas se espalham e uma área de alta concentração para uma área de baixa concentração.
Para poder alcançar uma imagem de qualidade, o modelo sempre inicia com uma imagem com muito ruído, e a partir da interpretação do texto enviado pelo usuário, a imagem irá iterando por etapas e sofrendo melhorias até chegar na sua versão final.
Uma vantagem é que esse é um dos modelos que tem um baixo custo, mesmo assim compete com outros modelos por conta de sua eficiência ao final do processo.
Além disso, o Stable Diffusion 3 conta com uma arquitetura de transformação de difusão multimodal. Em resumo, essa arquitetura atribui pesos para representações de imagem e linguagem a partir dos dados utilizados como treinamento, e depois utiliza esses pesos para uma imagem mais aderente à solicitação final do usuário, que seja nítida e com qualidade.

Dall-e:

Esse modelo utiliza a técnica de rede neural generativa para geração das imagens, transcrevendo textos imputados nos prompts dos usuários, e os transformando em imagens. Esse modelo tem uma base imensa de treinamento, que consiste em imagens com suas respectivas descrições.
Outro ponto essencial do Dall-e, é a utilização de GPT (generativa pre-trained transformer) em sua arquitetura. Esse modelo é responsável por não só consumir os prompts do usuário, mas também entender as nuances da linguagem, seu significado e sua importância no contexto solicitado. Isso é essencial para que a imagem saia com uma boa qualidade, até mesmo quando um prompt é curto, sem especificações, o GPT ajudará a criar uma imagem magnífica com o que foi lhe passado.
A geração das imagens do Dall-e funciona de maneira hierárquica. Baseado nos milhões de dados utilizados para o seu treinamento, a imagem é gerada de maneira topdown, ou seja, consome os aspectos essenciais da imagem e posteriormente refina os detalhes passados para o LLM.
Exemplo: Faça um leão andando pela Savana, com pelos dourados e dentes afiados.

No exemplo acima, o LLM irá primeiro focar na solução de geração do leão com a Savana, e isso não ocorreu por conta que foi o primeiro passo lido, mas porque é a parte essencial da imagem. Somente após ter essa parte definida, o modelo irá gerar os detalhes de “dentes afiados e pelos dourados”.

Diferenças entre modelos Dall-e e Stable Diffusion (SD):

Dentre os dois modelos citados até agora, temos algumas diferenças importantes para serem citadas.
Assim como seu coirmão de LLM, o Chat GPT, o Dall-e não tem um modelo open source (código aberto) para que seu usuário possa entender mais a fundo como ele funciona. Já o SD é um modelo totalmente open source, com total transparência do seu uso.
Outro ponto extremamente importante dentre suas diferenças, é o fato que o SD usa modelos de difusão estável, e o Dall-e usa os Transformers do GPT. Isso impacta na maneira em que cada um dos modelos irá gerar a imagem. O SD por exemplo, utiliza seu modelo de difusão para iniciar com um modelo totalmente aleatório e “ruidoso” e através de etapas, irá melhorando a imagem cada vez mais até chegar no resultado claro e com qualidade. Já o Dall-e, por meio dos seus transformers, utiliza uma base imensa de dados no seu treinamento que são representadas por imagens e suas descrições. A partir dessa base, o modelo utiliza o GPT para consumir o prompt do usuário, identificar os principais pontos essenciais da imagem e de maneira hierárquica, construir os pontos essenciais e depois refinar com os detalhes do usuário.
O uso das duas ferramentas, por mais que geram o “mesmo” resultado, são diferentes. Dall-e utiliza textos em prompts simples pelo usuário, já o SD utiliza linhas de código, ou seja, um conhecimento técnico é necessário para utilização do SD.
E por fim, o resultado de ambas as imagens trazem conceitos diferentes. No Dall-e, sua imaginação é o limite para a geração de imagens, já no SD, o resultado fica mais próximo de artes com técnicas avançadas de pintura, como realismo, cubismo ou pintura em óleo.

Midjourney:

Um dos modelos mais famosos, o midjourney tem algumas características interessantes no seu uso. Uma delas, é o fato que, diferentemente de SD, onde você precisa ter conhecimento técnico para utilizar, e Dall-e, que você acessa através de uma ferramenta, o uso do Midjourney é totalmente pela plataforma de comunicação Discord. Através dela, você gera os prompts da imagem e envia no chat para geração de imagem. Isso tem um custo atrelado, diferente das outras opções que ao menos trazem testes gratuitos.
Assim como o SD, o midjourney também funciona com método de difusão, ou seja, inicialmente com uma imagem repleta de ruídos e totalmente aleatória, e aos

poucos, faz com que a imagem sofra uma melhora na resolução chegando a um resultado claro e com qualidade.

Questão Letra B: Utilize diferentes técnicas de “Estilo Visual” e “Composição”, além de exemplos com *negative prompting*, para gerar 3 versões de imagem para cada proposição e avalie as diferenças entre os resultados (as imagens) e os prompts (as proposições).

Resposta Letra B:

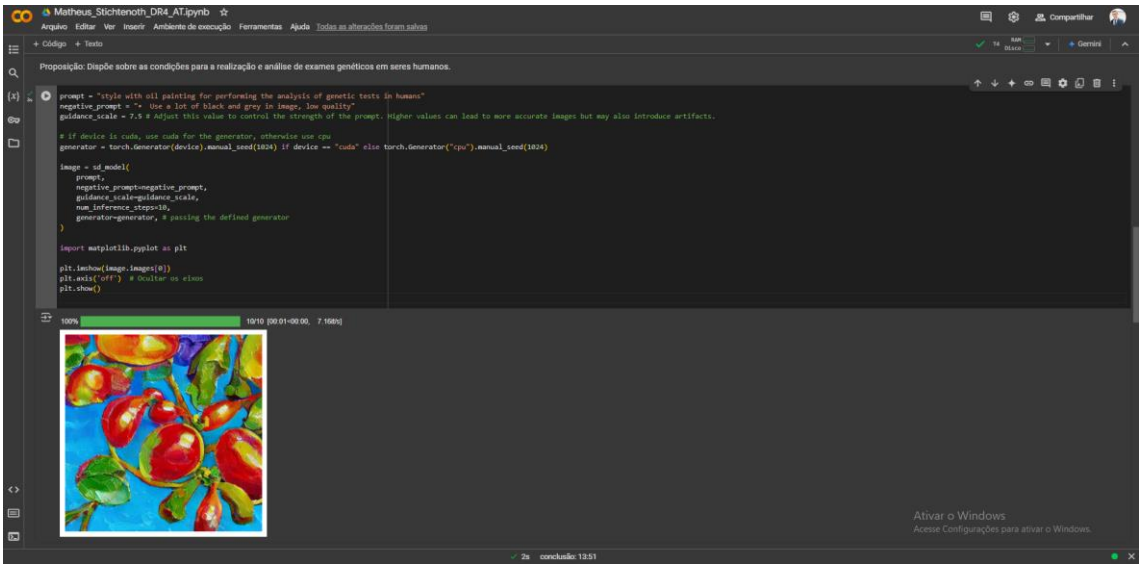
Ementas selecionadas:

- Dispõe sobre as condições para a realização e análise de exames genéticos em seres humanos.
- Institui o Prêmio Brasil de Incentivo à Pesquisa e à Aplicação de Conhecimentos e de Tecnologia para o Desenvolvimento Humano (Prêmio Brasil).

Para criação dos prompts, tive que resumir para que o LLM consegui entender a informação passada:

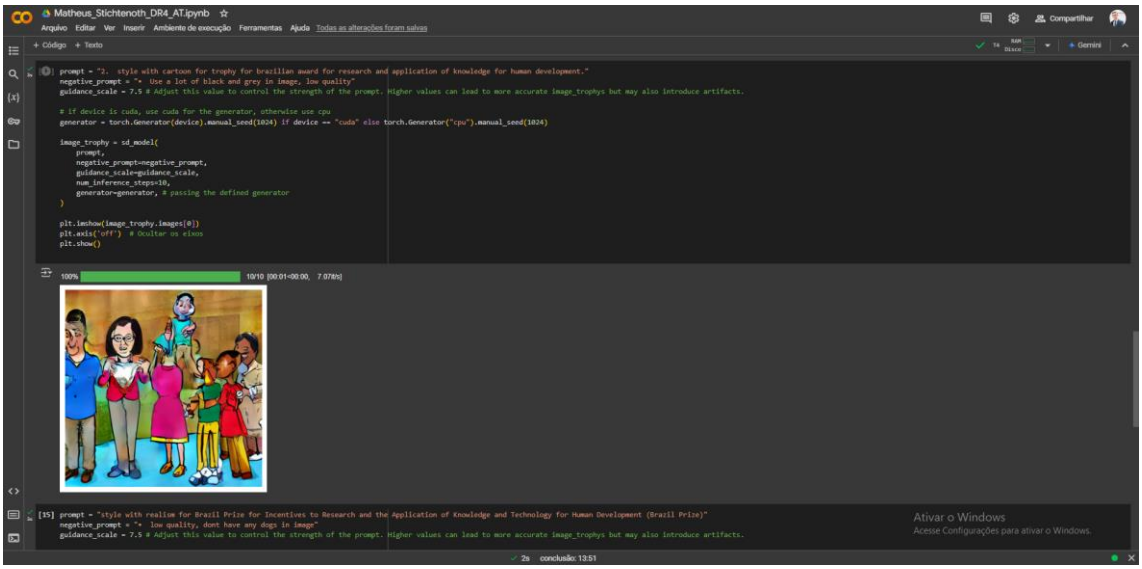
1. style with oil painting for performing the analysis of genetic tests in humans.
2. style with cartoon for trophy for brazilian award for research and application of knowledge for human development.
3. style with realism for Brazil Prize for Incentives to Research and the Application of Knowledge and Technology for Human Development (Brazil Prize)

Retorno do LLM:

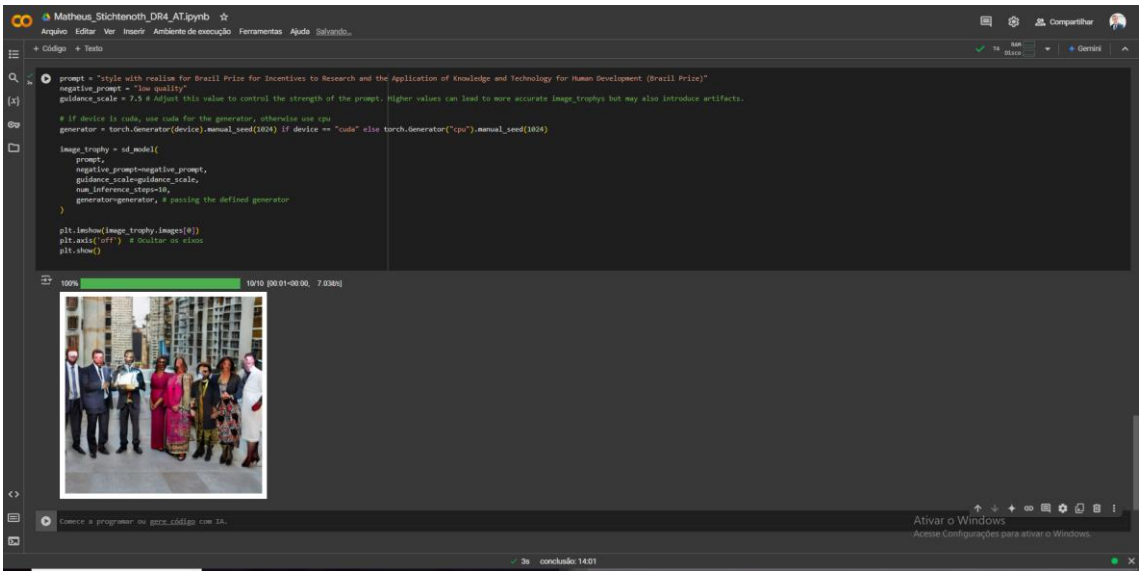
1. 

O resultado da primeira imagem não condiz com o prompt passado. A imagem remete muito à frutas. Isso pode ter ocorrido por conta do estilo da imagem solicitada, ou porquê o LLM não conseguiu captar o real objetivo

do prompt, e acabou desviando, causando quase que uma alucinação, por mais que fora utilizado uma escala de guia alta, que tende a gerar resultados mais próximos às solicitações.
O prompt negativo buscava afastar qualquer possibilidade de cores muito escuras no imagem.

2. 

O segundo prompt tinha como objetivo mostrar algo como um troféu para um prêmio brasileiro. O resultado é que o troféu não fica identificável na imagem. Imagino que o troféu seria o objeto triangular do personagem de roupas rosa e de óculos. A informação de que era algo que remetia ao Brasil, não teve diferenças no prompt com algo que marcasse que essa imagem era de um prêmio brasileiro.
Poucas cores escuras foram utilizadas, isso significa que o negative prompt teve sucesso nas instruções passadas à ele.

3. 

Utilizando um prompt um pouco diferente, mas com a mesma ideia da proposição anterior, foquei em um estilo realista para essa imagem, permitindo que as cores mais escuras fossem utilizadas.

O resultado aparenta estar mais aderente ao solicitado, sendo possível identificar que o troféu está com a terceira pessoa, da esquerda para a direita. Os rostos ainda seguem bem desconfigurados, mas imagino que isso seja limitações do LLM.