Machine Learning in Production
Midterm, Fall 2022

Christian Kaestner

Name: _____

Andrew ID: _____

Instructions:

- Not including this cover sheet, your exam should have **9** pages. Make sure you are not missing any pages.
- All questions in this midterm refer to the scenario on the first page. Answers are graded in the context of the scenario; **generic answers that do not relate to the scenario will not receive full credit.**
- The exam has a maximum score of **54** points. The point value of each problem is indicated. We allocated approximately one point per minute.
- We give an amount of space commensurate with what we expect you to need for each question. We use horizontal lines to suggest where to not use the full page. You may exceed those limits if it is clear where to find the rest of your answer. However, we strongly recommend writing concise, careful answers; short and specific is much better than long, vague, or rambling.
- **Do NOT write anything that you want us to grade on the back of pages.** We will scan the exam and will not look at the back sides.
- This is a **closed book exam**; no books or electronics allowed. You may refer to 6 sheets of notes (handwritten or typed, both sides).

Scenario: Interactive Art

Image generation techniques currently explode in machine learning research. First *DALL*E* made a splash, then *Stable Diffusion* was released publicly in mid-August and has triggered a huge amount of excitement and experimentation. These techniques can generate images from a text prompt and can imitate many artists. For example, the image on this page was created by DALLE 2 with the prompt *"van gogh painting of students in a large classroom taking an exam."* More recent experimentation focuses on videos and 3D scenes generated with text prompts.

With two friends, you have an idea for a business: You want to sell accessible interactive art to museums and other institutions (e.g., airports, city hall, offices). You think of a large display on which users could sketch an image with their fingers and speak a voice prompt of what they try to see and the system will turn this into a short video animation in the style of an artist that is relevant to the customer (e.g., Andy Warhol in the Andy Warhol museum or PIT airport, Van Gogh in the Amsterdam city hall). Users can also sketch and describe modifications to existing images the customer may provide. You expect customers with minimal technical expertise who want to buy the installation as an easy to use package (hardware, software and support). After some initial market research you expect that you can charge a few early customers $50,000-$200,000 for the initial installation and $500-$1000 per month service fee.

You plan to build on recent advances with *Stable Diffusion*. The model training for the core image generation model was very expensive (trained on a large cluster over a month, estimated at $600,000 for just the final training run) but the model can now be used freely under an open license. The model itself takes 36 GiB disk space. Model inference is also expensive: With a standard graphics card, generating a single image will take about 30 seconds. You have some ideas for how to create videos much faster than computing every frame separately.

You do not have a lot of startup capital. Your team is small and you do not expect to be able to hire additional technical team members before deploying the first products. Your existing team of three has sufficient data science expertise and is familiar with the research on image generation, and two team members feel comfortable with software engineering. You have some high-end hardware to get started and often rely on cloud computing resources, but you won't be able to afford to train your own image generation models entirely from scratch.

Question 1: Goals and Telemetry [16 points]

To understand the requirements of the overall product better, you first analyze various goals from different perspectives. For two of these goals you also already plan how to eventually measure how well your solution meets the goal.

(a) **User goal from the perspective of the customer (museum, airport, …)**

Goal:

Proposed measure:

Data to collect (existing or additional):

Operationalization:

(b) **User goal from the perspective of a person interacting with the system (general public, museum visitor, …)**

Goal:

(c) **Model goal of the model generating short video from sketch + text prompt**

Goal:

Proposed measure:

Data to collect (existing or additional):

Operationalization:

(writing below this line is allowed but discouraged)

Question 2: Trade Offs [14 points]

You have already determined that the computational cost for image and video generation is too high to run on an embedded device in a smart display. You are trying to decide whether (a) to deploy the Interactive Art installation entirely onsite at the customer or (b) to move pretty much all computations into the cloud. You estimate that you could use a high-end PC with a cutting-edge GPU that connects to the display onsite (extra cost of around $6000). You are not sure yet what setup to use and try to collect more information to make a decision.

(a) [4 points] Identify two *system qualities* and two *model qualities* that you consider to be of high importance in the scenario. Give a 1-sentence justification each why they are important in the scenario.

System quality 1:

System quality 2:

Model quality 1:

Model quality 2:

(b) [6 points] Considering how the two deployment options support or hinder the four qualities, where would you recommend to generate the images and videos? Briefly justify your answer by explicitly referring to the three qualities. Try to stay generally within the realism of the scenario. If you need more information, assume you have done additional research and simply state the hypothetical finding on the next page.

☐ On-site PC is better ☐ Cloud is better

Justification:

(c) [4 points] Under which conditions (e.g., alternative requirements prioritization, alternative hardware capabilities, …) would you recommend the opposite deployment option?

Optional. Additional hypothetical research findings informing your answers:

Question 3: Model Quality [10 points]

(a) [6 points] As part of the system, you want to accept voice prompts. Since this is standard technology, you plan to use an existing model, rather than training your own. Today, you are trying to evaluate whether AWS's speech-to-text model API is suitable for your task. Their model card promises high accuracy in different situations, but you are worried whether their test data is representative of your use case. Briefly describe how you could evaluate the model yourself using the idea of slicing and simulation. Ensure that your answer demonstrates an understanding of the concepts and relates to the scenario.

Slicing:

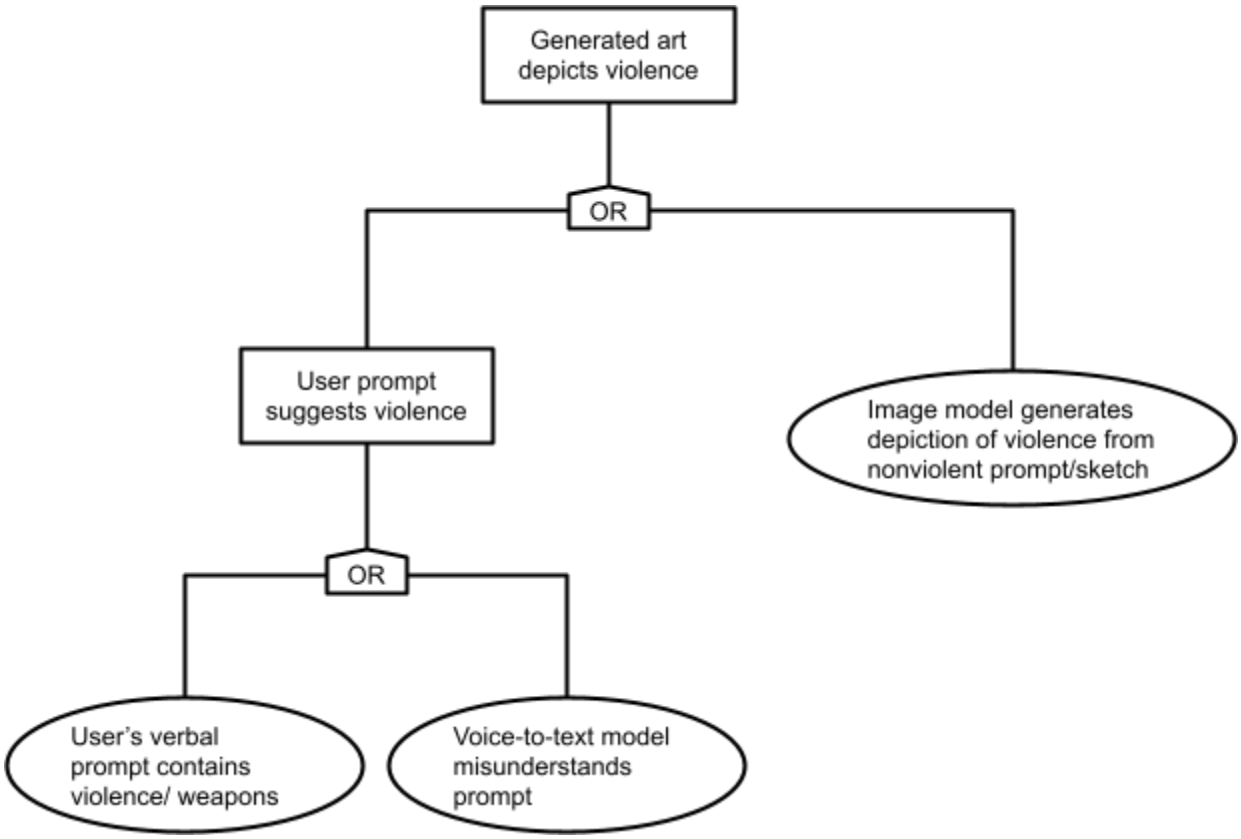
Simulation:

(b) [4 points] You plan to occasionally update the image/video generation models. Briefly discuss whether *canary releases* would be useful in this scenario. Ensure that your answer demonstrates an understanding of the concept and relates to the scenario.

Question 4: Mitigating Mistakes [14 points]

You and your customers are naturally concerned about potential abuse of the system if it operates in public spaces. Different customers are likely concerned about different issues, but you expect that many want to avoid profanity, nudity, violence, and overt political messages. For now, you focus on the following requirement for an early customer: *The generated art that is shown publicly should not depict violence (weapons, war, fist fights, …).*

You started with a simple fault tree to analyze possible causes and mitigations:



(a) [4 points] Describe a mitigation strategy to reduce the risk of generating art that depicts violence that *does not rely on human involvement*. The mitigation should be at the system level, outside of the ML component (i.e., not just "train a more accurate model"). Your answer must make clear how the mitigation reduces the likelihood of the requirement violation.

(b) [2 points] State one *software specification* that relates to the implementation of your non-human mitigation strategy.

(c) [2 points] Update the fault tree above to reflect your mitigation (you can add to or cross out parts of the existing fault tree).

(d) [4 points] Describe an additional human-in-the-loop mitigation to reduce the risk of generating art that depicts violence. Explicitly refer to the concepts of automate, prompt, augment/annotate/organize and explain how the mitigation reduces the likelihood of the requirement violation. (You do not need to update the fault tree.)

(e) [2 points] State one *environmental assumption* that is necessary for your human-in-the-loop mitigation to work.

(writing below this line is allowed but discouraged)