Machine Learning in Production / AI Engineering Midterm, Spring 2024

Claire Le Goues and Christian Kaestner

Name: ________________________________

Andrew ID: ________________________________

Instructions:

- Not including this cover sheet, your exam should have **8** pages. Make sure you are not missing any pages. You may detach the last page.
- All questions in this midterm refer to the scenario on Page 8. 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 designed the exam anticipating approximately one minute per point.
- **Please write legibly**. We are unlikely to be able to grade your solution if we can't read it.
- 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 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).

Question 1: Goals and Telemetry [10 points]

All questions in this exam relate to the scenario on the last page. You may detach the last page if you like. Your first task is to explore and document goals for the TranquilAI product and identify how you can measure success.

(a) [4 points] State an *organizational goal* for Apple and a *key performance indicator* that can approximate the contribution of TranquilAI to that goal with relatively low latency.

Organizational goal:

Key performance indicator:

(b) [6 points] You plan to evaluate how the model for *Relaxation and Discomfort Sensing* does *in production*. In particular, you would like to see how often the model detects the user being distracted when they actually are not. Design a measure and suggest what data to collect and how to operationalize the measure with telemetry. The measure can be an approximation, but must be plausible within the realism of the scenario.

Measure:

Data to collect (what and how):

Operationalization:

Question 2: Trade-offs [15 points]

You are considering how to deploy the three models of the TranquilAI product (see the scenario on the last page), whether on the VisionPro hardware, in the cloud, on a laptop, or with a dedicated hardware device sold with the product (with a GPU).

(a) [5 points] Identify and roughly rank two qualities that are important for the decision in this scenario and one quality of little importance (no measure required for any of them). Provide a brief justification of why they are important or not important:

Quality 1 (most important):

Quality 2 (second most important)**:**

Quality 3 (low importance):

Justification:

(writing below this line is allowed but discouraged)

(b) [6 points] Make a recommendation with a brief justification of how/where to deploy the models for *Relaxation and Discomfort Sensing*, considering the tradeoffs between the qualities. Refer explicitly to the important qualities identified previously and underline them in your text. Your answer must relate to the scenario. If you are missing information to make that decision, describe what information you would need and how you would make a recommendation with it.

(c) [4 points] Do you recommend the same deployment approach for the *Video Generation* model? Briefly justify why or why not. Your answer must relate to the scenario and the identified important qualities. Answers without a justification will not receive credit.

□ *same approach* □ *different approach*

Justification:

Question 3: Model and Data Quality [12 points]

(a) [4 points] Early experiments with machine learning for *Relaxation and Discomfort Sensing* (see scenario on the last page) create results that seem too good to be true. Name one *pitfall* in model accuracy evaluation that could cause too positive accuracy evaluations in this scenario and describe how you would detect or avoid the problem. Your answer must demonstrate an understanding of the pitfall and relate to the scenario.

(b) [4 points] You consider whether the evaluation can be improved with the idea of *slicing.* Briefly discuss whether or not slicing would be a suitable approach here and why. Your answer must demonstrate an understanding of the benefits or drawbacks of slicing and relate to the scenario.

(c) [4 points] Provide a plausible concrete example of *data drift* in the scenario (i.e., changes in data distributions, not decision boundaries) that may degrade the accuracy of your model for *Relaxation and Distraction Sensing* in production over time.

Question 4: Risks and Mitigation [13 points]

To plan for mistakes you try to better understand the requirements and risks of the product (see the scenario on the last page). For the following, you focus on the following safety requirement:

> *“When the user experiences motion sickness or disorientation (e.g., nausea, dizziness), TranquilAI pauses all video and audio within 2 seconds.”*

(a) [3 points] Classify the following parts of the TranquilAI product into world and machine entities (in the world vs machine sense from the reading and lecture):

- Motion sickness of the user □ *world entity* □ *machine entity*
- The heart rate sensor in the smartwatch □ *world entity* □ *machine entity*
- The deep neural network processing sensor inputs □ *world entity* □ *machine entity*
- Historic sensor readings stored in a database □ *world entity* □ *machine entity*
- Video content generated by a generative model □ *world entity* □ *machine entity*
- The wifi connection between the smartwatch and the VisionPro headset □ *world entity* □ *machine entity*

(b) [2 points] State one **software specification** that is necessary for the system to satisfy the above requirement.

(c) [2 points] State one **environmental assumption** that is necessary for the system to satisfy the above requirement.

(writing below this line is allowed but discouraged)

(d) [6 points] If the model incorrectly predicts that a user is fine while they experience motion sickness, the requirement above may be violated. This would correspond to paths in a fault tree (you do not need to draw the tree). Describe a *mitigation* to make it less likely that the requirement will be violated even if the model prediction is wrong. *The mitigation should be at the system level, outside of the ML component* (i.e., not just "train a more accurate model" or “use an ensemble model”). In addition, check the box corresponding to whether your mitigation would eliminate a basic event from the fault tree to add another basic event.

Update to a fault tree (check one, no further explanation needed):

□ *eliminate basic event* □ *add basic event*

Mitigation description:

Question 5: Git [4 points]

Your team exchanges some ideas, mostly text documents, with your academic collaborators through private GitHub repositories. How would you explain to a colleague the purpose of *amending* a commit and when it is appropriate. Your answer must demonstrate an understanding of the concept and must convey at least one benefit.

Scenario: TranquilAI

You have joined a team at Apple that is developing an innovative in-house application for the newly-released Vision Pro mixed-reality 3D headset. Your product, code-named TranquilAI, will provide adaptive, immersive guided meditation experiences, leveraging recent advances in generative AI for video and sound. Imagine meditating on a serene, beautiful beach, from the comfort of your living room.

TranquilAI will stand out in a crowded field because the experiences will be *adaptive* to biometric data from a user's Apple Watch. Is the user's pulse rate a little elevated? Maybe she'd be calmed by a beautiful mountain sunset. Is she restless, with her watch indicating movement? The session guidance can adapt by focusing on deep breathing exercises.

TranquilAI therefore has several core pieces: (1) **Video Generation:** AI-generated audio and visual 3D environments. (2) **Narrative Generation:** Guided narrated meditation content that provides a set of scripts varying in focus, such as breathing techniques or visualization exercises (3) **Relaxation and Discomfort Sensing:** Biometric feedback analysis that receives real-time data on the user's pulse rate and movement from the user's smartwatch, along with novel algorithms that interpret this data to assess the user's level of relaxation, distraction, or potential physical discomfort (e.g., dizziness from the headset). Importantly, the product is **adaptive** in that it (a) adjusts content generation (components 1 and 2) based on real-time biometric feedback (component 3) and (b) learns which environments and types of meditations work best for the user, personalizing the experience over time.

The system is also expected to live up to Apple's reputation for both simple and intuitive UI, as well as rigorous privacy protection, especially for sensitive biometric data.

Although Apple is in good financial shape overall and is invested in the success of the Vision Pro platform, no company is immune to the expense and limited availability of GPUs – you compete with many other teams within Apple over resources for training and validation. Your team works with existing foundation models for text, audio, and video generation from other groups at Apple, but is responsible for interpreting biometric feedback and creating the overall adaptive meditation experience. Your core team is small, with 3 experienced data scientists and 2 software engineers with experience in embedded systems. You have access to several meditation teachers in the local area who are excited about the project and have connections to two Psychology research labs at East Coast Universities that study meditation academically. Apple does not hire lightly, so you do not anticipate being able to recruit additional team members, and the timeline is tight: You want to release a product within the next 6 months.