New interview guide

Notes:

- **When a practitioner mentions a quality issue, always ask how they prevent the problem from happening**.

Guide

Introduction

▼ Short introduction of the interviewers.

▼ Description of the interview

▼ Goal of the interview

- **What**: Develop a catalog of quality issues in Machine Learning Software Systems.
- We are interested in issues you have encountered while building Machine Learning Software Systems (MLSSs) that affected the quality of the developed system. MLSSs are any software system with a ML component (component that relies on ML for its functionality)
- **What is a quality issue**: Any issue that does not affect the functionality of a system, but only its serving quality. For example, a recommender system whose predictions are accurate but not explainable has quality issues, but not functional issues.
- **Quality aspects of ML**: robustness, scalability, explainability, model complexity, resource demand, etc.

- We will ask you approximately 20 questions

▼ Setting up the interview

1. Ask for permission to record the interview. Explain to the interviewee that it is our intention to release an anonymized version of the interview transcript

publicly.

2. Some background information
    a. Current position
    b. Experience (general/specific to ML)

Body of the interview

Present the structure of the interview

▼ [Q1] A general and open-ended question to start the interview:

- What are the main quality issues that you have encountered with your data, model or system so far?

▼ [Q2] Data collection questions

- [Q2.1] Do you use any of the following data collection technique?

    ▼ Data collectors
    Anyone manually creating training data for a training algorithm (e.g. an employee filling reports with information ingested by a ML model, a radiologist labeling x-ray scans, etc.).

    1. Have you experienced any data quality issue with this process?
    2. How do you verify the quality of the data collected and how do you ensure its quality?

    ▼ External data
    For example: public datasets, third-party API or web-scrapped data.

    1. Have you experienced any data quality issue with this process?
    2. How do you verify the quality of the data collected and how do you ensure its quality?

    ▼ Data generated by another system (ML-based or not)
    For example: (1) a model using past sales to predict future sales of a product (where past sales are automatically saved by a system) or (2) a model using

the prediction of another model predicting the weather to predict the sales of a food product.

1. Have you experienced any data quality issue with this process?
2. How do you verify the quality of the data collected and how do you ensure its quality?

▼ [Q3] Data preparation questions (data cleaning + data transformation)

1. [Q3.1] Which data types have you worked with?
    a. Have you encountered quality issues with these data types? What were they?
2. [Q3.2] Have you ever measured the quality of your data and/or tried to improve it?
    ▼ If yes
        - How?
        - Do you have tools/frameworks that help you clean your data?
3. [Q3.3] What are the issues you repetitively encounter when preparing data for ML?
    a. Why and how do these problems happen?
    b. How do you handle these issues?
4. [Q3.4] Is there any other data quality issue we missed that you consider relevant?
    ▼ If yes

        How do you handle the issue?

▼ [Q4] Model evaluation questions

1. [Q4.1] How do you evaluate the quality of models? As a reminder, quality is not only defined by ML performance, but also by other aspects, such as explainability, robustness, scalability, etc.
    - Do you have tools/frameworks that help you with that?

2. [Q4.2] Have you used existing benchmark models for quality aspects to evaluate your model?
3. [Q4.3] Have you ever assessed the quality of an ML model's predictions with the users of your system?

    ▼ If yes

    How have you proceeded?

4. [Q4.4] Have you ever assessed the quality of an ML model's predictions with subject matter experts (SME)? Subject matter experts are people with a good understanding of the problem that must be modeled by a ML model.

    ▼ If yes

    How have you proceeded?

5. [Q4.5] Have you encountered any other quality issues during the evaluation of your models?

    ▼ If yes

    How do you handle the issue?

▼ [Q5] MLSS deployment questions

1. [Q5.1] How (manually vs. automatically) and where are your models deployed?
2. [Q5.2] What are the challenges you have encountered during the deployment of a MLSS?
3. [Q5.3] Did you ever have a model that performed well locally but poorly once deployed?
    a. [Q5.3.a] What caused this problem?
    b. [Q5.3.b] How did you handle the problem and what are the measures taken to prevent it from happening in the future?
4. [Q5.4] Have you encountered any other quality issues with your model or system during the deployment phase?

    ▼ If yes

    How do you handle the issue?

▼ [Q6] MLSS maintenance questions

1. [Q6.1] How do you ensure that the quality of a MLSS does not decrease over time?
2. [Q6.2] Have you encountered issues with data (i.e. the data sources or the data) during the maintenance of a MLSS?
   e.g. unreliable data sources, concept drift, etc.

   ▼ If yes

   1. What are the issues?
   2. Do you have mechanisms to prevent the issue from happening again? If yes, what are they?
3. [Q6.3] Have you encountered issues with the model during the maintenance of a MLSS?
   e.g. model staleness, unreliable ML performance between re-trainings

   ▼ If yes

   1. What are the issues?
   2. Do you have mechanisms to prevent the issue from happening again? If yes, what are they?
4. [Q6.4] Have you had other issues regarding the maintenance of your models or system?

   ▼ If yes

   How do you handle the issue?

▼ [Q7] Quality measures of ML models questions

- Did you ever had issues with one of the following quality aspect:

  ▼ Fairness

  - Definition: People of different groups (i.e. race, ethnic origin, religion, gender, sexual orientation, disability or any other personal condition) should not be treated in a discriminative way.
  - Example(s) of MLSSs with quality issues:

    - A face recognition system that does not detect people of dark skin color.
    - A job prediction system that associate some jobs with sex (e.g. nurse).

▼ Robustness

- Definition: Robustness deals with situations where the model is out of its normal operational conditions. Robustness characterizes the resilience of the model.
- Example(s) of MLSSs with quality issues:
    - An auto-pilot system that can not detect a stop sign if it is held by a human (i.e. a road worker) instead of a pole.

▼ Explainability

- Definition: "*AI in which humans can understand the decisions or predictions made by the AI*" (taken from here).
- Example(s) of MLSSs with quality issues:
    - A loaning system that can not explain why a person has been given a poor deal.

▼ Scalability

- Definition: The measure of a system's ability to increase or decrease in capacity or functionalities based on external factors.
- Example(s) of MLSSs with quality issues:
    - A MLSS that can not efficiently leverage a lot of computing power to increase its performances.

▼ Privacy

- Definition (what is meant in the context of ML): to avoid the leakage of data or models.
- Example(s) of MLSSs with quality issues:
    - Data: An auto-complete system that leaks confidential information.

- Model: A MLSS that shares too much information regarding its prediction which enables an adversary to extract a model's parameters of a MLSS by observing its predictions (Tramer et al.).

▼ Security

▼ Archive

1. [Q7.1] Have you ever faced problems related to the scalability of trained models (e.g. scalability regarding the number of machines it is deployed onto)?
    a. [Q7.1.a] What kind of scalability issues?
    b. [Q7.1.b] How does your team currently handle these issues?
2. [Q7.2] Is robustness a significant quality issue when building ML models? Robustness deals with situations where the model is out of its normal operational conditions. Robustness characterizes the resilience of the model.
    a. [Q7.2.a] How do/did you evaluate robustness?
    b. [Q7.2.b] How does your team currently handle robustness issues?
3. [Q7.3] Have you ever investigated the explainability of your trained models? (trying to explain the final decision of the model)
    a. [Q7.3.a] How? Which measurement did you use?
4. [Q7.4] Is there any other quality issue in ML systems that you have experienced and that we did not inquire about in this interview?
    ▼ If yes

    How do you handle the issue?

Conclusion

▼ Two closing questions

[Q9] In your opinion, what is the most pressing quality issue researchers should try to solve?

[Q10] Do you have any other comments about the quality of ML systems?